“They Edited Out her Nip Nops”: Linguistic Innovation as Textual Censorship Avoidance on TikTok

Kendra Calhoun
University of Illinois, Urbana-Champaign

Alexia Fawcett
University of California, Santa Barbara

Abstract

In response to content moderation that disproportionately censors discourse by and about marginalized users, content creators on the video sharing platform TikTok have developed a linguistic repertoire whose communicative effect is akin to that of an avoidance register. Creators manipulate sound, morphology, meaning, orthography, and gesture to circumvent lexical items that may be censored based on notions of “appropriateness” articulated in the platform’s community guidelines. The strategies they use parallel documented forms of language play online and offline, as well as censorship avoidance on other social media platforms. These strategies are used most frequently on words related to contested ideas such as race, gender, and sex(uality). However, because of the memetic culture of TikTok, the practice of linguistic self-censorship has expanded to contexts where content is at little risk of top-down censorship, instead functioning to reflect creativity, make social commentary, and index sociopolitical alignment.

Introduction

The video-sharing app TikTok reported one billion monthly active users in September 2021 (TikTok, 2021). This milestone was fueled, in part, by massive uptake of the app during global COVID-19 lockdowns. While still popular for memes, dance challenges, and lip synching, TikTok content now ranges from comedic performance and everyday moments to “life hacks” and social critique (Schellewald, 2021). Like on other social media, a significant amount of TikTok content focuses on the self (Bhandari & Bimo, 2020), and among creators from structurally marginalized groups – e.g., people of color, people who are queer, trans, disabled, and/or undocumented – videos about their everyday experiences often reflect the reality of inequality in U.S. society. These topics are frequently seen as contentious by those they call out, who attempt to quell critical discourse about themselves by reframing it as offensive or violating some set of shared values (e.g., Bucholtz, 2019).

Social media users can appeal to the supposedly shared values of platforms’ community guidelines. These guidelines, varying in scope and enforcement, dictate what is “appropriate” content and behavior. They are rarely community-driven, instead implemented in a top-down fashion: platform operators develop them as a condition of use to protect themselves legally, and they may be changed as platform operators see fit (e.g., Diaz & Hecht-Felella, 2021). Guidelines can be enforced by humans, such as social media users reporting individual posts for review by content moderators. They can also be enforced algorithmically by computers trained on such guidelines (e.g., Roberts, 2019), and this algorithmic enforcement can be built into a platform's features. For example, when TikTok first released auto-generated speech-to-text captions, spoken curse words were omitted from the text. Whether through human or algorithmic detection, posts deemed potentially offensive may be removed completely or shadowbanned (restricted from content recommendations) and the accounts that upload them may be suspended. When platforms attempt to exert algorithmic control over users, however, users innovate ways to maintain control of their content (e.g., Gerrard, 2018). To avoid what some call being "guidelined," TikTok users now creatively circumvent lexical items that may be deemed violations of community guidelines. Because of the mimetic nature of language on TikTok (Zulli & Zulli, 2020), this practice has also expanded to contexts where potential censorship is a minimal threat. Although all social media platforms implement some form of content moderation, TikTok has garnered significant public attention due to the high number of users and the platform's popularity as a source of income for creators. There are hundreds of millions of people on TikTok to see and discuss when content is censored and an ever-growing number of people whose economic livelihoods are at stake should their content be censored.

In this article, we examine the linguistic self-censorship practices of TikTok creators and the social and technological factors that shape them. For the purposes of our analysis in this context, we define *linguistic self-censorship* as:

> any instance of a social media creator intentionally changing their linguistic practices to avoid using a specific word or phrase, either because of potential risk for actual censorship through algorithmic enforcement of community guidelines or as a form of mimetic language play.[1]

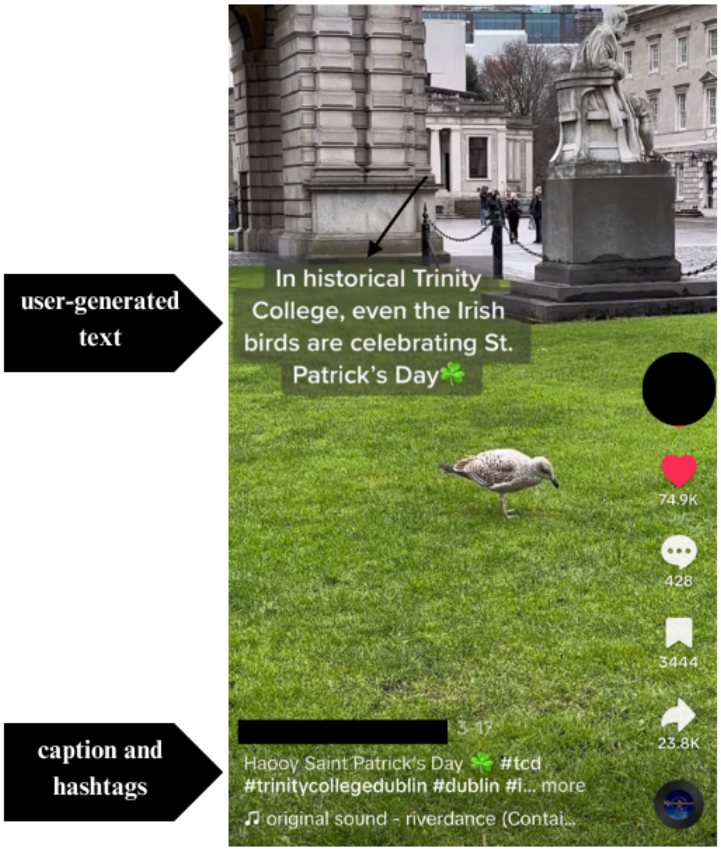
Specifically, we analyze English-language data for (1) the types of words creators choose to censor, (2) the linguistic processes they use to create censored forms, (3) how TikTok's platform features afford these linguistic changes, and (4) the social and interactional function of self-censorship beyond minimizing one's chances of top-down censorship. After a brief description of TikTok below, we discuss the theoretical context for our analysis; we then describe our data and methods, followed by an analysis of select examples from the data set. In the discussion, we return to a consideration of how this linguistic phenomenon reflects how social biases and digital technologies interact, impacting social media users from marginalized groups; how self-censorship on TikTok compares to other forms of language play; and the contextual nature of both the production and interpretation of these forms.

TikTok

While TikTok's "Following" feed allows users to see content exclusively from creators they follow, most users primarily engage with videos via their "For You Page" (FYP). TikTok's algorithm initially populates the FYP with viral videos and topics of interest a user selects when joining the app and refines content based on their viewing and engagement practices over time.

For example, both authors see many videos about language, but we also see different "sides" of TikTok based on our respective identities, interests, and locations. Calhoun sees "Black TikTok," digital art, service workers, and content about the Midwestern U.S., whereas Fawcett sees pottery, curly hair care, miniatures, and content about the Western U.S.

TikTok users who create content are encouraged to participate in trends and imitation through the use of replicable "sounds" – audio from one video that can be used in another – and the viral success of specific styles of TikToks that garner engagement in the form of likes, comments, and shares (Abidin, 2021; Zulli & Zulli, 2020). In addition to audio, content creators have text, image, and various special effects at their disposal to participate in TikTok's memetic culture. Importantly, the suite of recording and editing features that makes TikTok's audiovisual content distinct also makes TikTok users' linguistic self-censorship practices stand out from those on other platforms (though they are not entirely novel). TikTok's text-based features include user-generated text that appears within the video, video description (caption), and hashtags (Figure 1), as well as comments and automated speech transcription (Figure 2) and a text-to-speech function.


In historical Trinity College, even the Irish birds are celebrating St. Patrick's Day☘

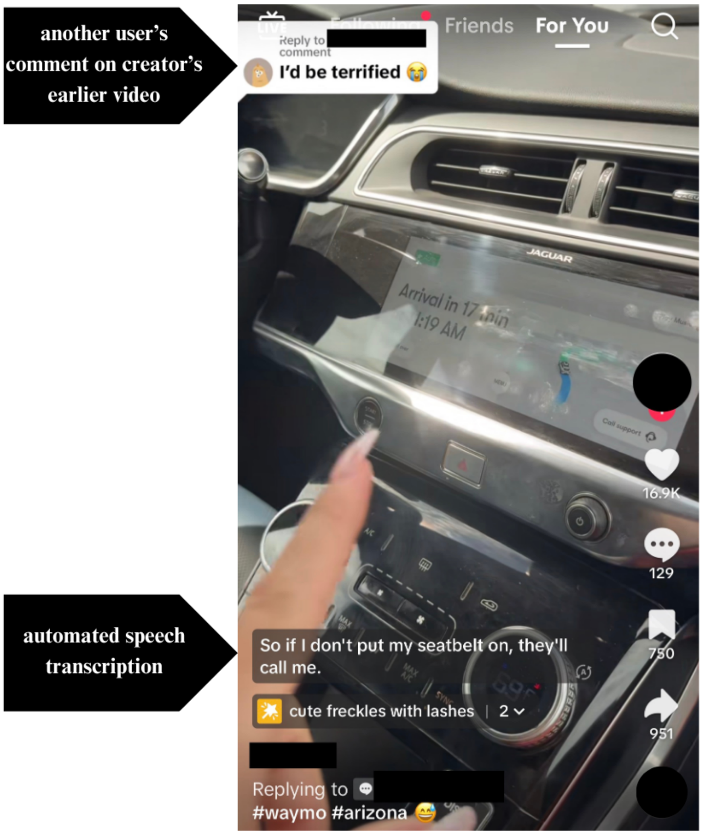
[redacted]

[redacted] 3-17

Haooy Saint Patrick's Day ☘ #tcd #trinitycollegedublin #dublin #i... more

♫ original sound - riverdance (Contai...

Figure 1. TikTok with user-generated text, video description, and hashtags


another user's comment on creator's earlier video

automated speech transcription

Reply to [redacted] comment

I'd be terrified 😭

So if I don't put my seatbelt on, they'll call me.

cute freckles with lashes | 2

[redacted]

Replying to [redacted]

#waymo #arizona 😅

Figure 2. TikTok with automated speech transcription, responding to a comment

Linguistic creativity features prominently on TikTok, often taking the form of humor and language play. For instance, a 2020 trend was to replace *pandemic* with various words starting with 'p' (e.g., *pandemonium*), and users continually one-upped each other with more outlandish words, including nonce words (e.g., *pandemilovato*). Alongside widespread trends, different TikTok communities use the audio, image, and text features of the platform to engage in in-group linguistic creativity that creates memeable content. TikTok users innovate discourse practices that are specific to the available technologies and interactional norms of the platform and digital community, a phenomenon that occurs in virtually all forms of digitally mediated communication (e.g., Calhoun, 2019).

According to TikTok's website, there are 12 main areas for content moderation, including expected topics such as violence and nudity (see Appendix). However, TikTok employs a black box system of "visibility moderation," meaning the structure of the system that determines which content and creators are promoted or suppressed is not made clear to users (Zeng & Valdovinos Kaye, 2022). Creators must therefore make choices about their content based on algorithmic folk theories, "intuitive, informal theories that individuals develop to explain the outcomes, effects, or consequences of technological systems, which guide reactions to and behavior towards said systems" (DeVito et al., 2017, p. 3165). Some words that creators censor are clear potential violations, such as *suicide* vis-á-vis guidelines on "suicide, self-harm, and violent acts"; however, context is crucial to determine whether a word is used in a way that violates the platform's rules. *Suicide* will inevitably be used by an account dedicated to suicide prevention, but automated systems do not distinguish that from content which promotes the act. Likewise, algorithmic detection may conflate talk about hate speech or ideologies with hate speech itself – erroneously flagging, for instance, a post in which someone discusses Nazi ideologies to point out their factual inaccuracies. Without context, an algorithmic system also cannot detect the difference between in-group uses of reclaimed words and derogatory out-group uses.

TikTok's official description of its community guidelines enforcement paints a picture of an egalitarian platform that proactively protects its users (see Gillespie, 2010). The website states:

> Our Community Guidelines apply to everyone and everything on TikTok. We proactively enforce them using a mix of technology and human moderation and aim to do so before people report potentially violative content to us…We will remove any content – including video, audio, livestream, images, comments, links, or other text – that violates our Community Guidelines. (TikTok, 2022)

However, tech news and popular media outlets have reported on accusations of TikTok disproportionately censoring and suppressing the content of creators from marginalized groups (e.g., Biddle et al., 2020). When TikTok leadership has admitted there is a problem, it has been framed as a fixable flaw in the system (Ohlheiser, 2021), but the widespread practice of self-censorship among TikTok creators suggests that many see it as bias embedded in the platform's current structure that must be navigated strategically.

Theoretical Context

As a linguistic practice, self-censorship on TikTok is a form of creativity and language play, which are not unique to digital media. Creators' intentional manipulation of language on TikTok has received significant attention, but language play for self-censorship online has existed as long as content moderation has been a policy. The digital infrastructure of online spaces and the sociopolitical context in which platforms, their creators, and their users coexist shape how self-censorship manifests on any given platform and the type of attention it receives. In this section, we situate self-censorship on TikTok historically and theoretically within these realities.

Language Creativity and Play

Beyond the baseline creativity needed for everyday social interaction, language can be used in ways that flout rules of standard use. Sherzer and Webster (2015, p. 1) describe this practice of "speech play" as:

> the playful manipulation of elements and components of language, in relation to one another, in relation to the social and cultural contexts of language use, and against the

> backdrop of other verbal possibilities in which speech play is not foregrounded. The elements manipulated can be at any level of language, from sound patterns to syntax, semantics, and discourse; they can include the various languages used in multilingual situations and relations between verbal and nonverbal communication. Speech play can be conscious or unconscious, noticed or not noticed, purposeful or nonpurposeful, and humorous or serious. Nonetheless, given the focus on manipulation, speech play typically involves a degree of selection and consciousness beyond that of ordinary language use.

Speech play, or more broadly, language play, is a response to a stimulus – be it a desire for social belonging or the imposition of hegemonic values by a dominant authority – and ranges in the "playfulness" of context. One end of the spectrum is game-like play such as puns, tongue twisters, rhymes, idioms, and linguistic parody (e.g., Norrick, 2017). The other end of the spectrum is "argot" (Valdman, 2000) or "anti-languages" (Halliday, 1976), coded language created to avoid detection and therefore sociopolitical and/or legal persecution. For example, Polari (Baker, 2002) was used by queer people in the UK for this purpose and Verlan (Lefkowitz, 1989) by lower-class residents of suburban Paris. In all contexts, language play demonstrates speakers' knowledge of language structure, and it can index identity and belonging to a particular community of practice (e.g., Basso, 1979; Eckert, 2006). Structural features of a particular play form may spread and usage evolve over time, but those who use an "anti-language" recognize a shared identity or worldview with interlocutors who do the same.

Language Play Online

Studies of linguistic creativity and play in online contexts importantly illuminate the interaction between language and technology. Gibbs et al. (2015) theorize this interaction through the concept of "platform vernaculars," asserting that while social media discourse, broadly, is a valid category of analysis, "each social media platform comes to have its own unique combination of styles, grammars, and logics" that are "of the people" (p. 257). Platform vernaculars

> can share many elements, and the vocabulary and grammars of vernaculars migrate between social media platforms as new practices and features from one platform are appropriated for use on others. [...] However, every platform has a vernacular specific to it that has developed over time, through design, appropriation, and use.

In other words, the linguistic features and practices used for language play on one platform likely occur on other platforms, but the manner in which and the purpose(s) for which they are used are platform specific. Platform vernaculars can also be analyzed as discourse genres, as theorized in literary studies, linguistics, and anthropology. Individual utterances can be analyzed as standalone texts, but they are also analyzable in terms of their structural, contextual, and rhetorical similarities to other utterances (Briggs & Bauman, 1992). Considering this, the linguistic processes used for the relatively new platform vernacular of self-censorship on TikTok can be understood as building on other creative platform vernaculars (e.g., fictional tumblr chats; Vásquez, 2019), and on earlier forms of self-censorship online – all while having unique features that warrant study in their own right.

Like their offline predecessors, language users online have developed language play as a form of resistance to top-down power structures. One early example is "leetspeak," whose name comes from *leet* or *1337,* short for *elite* (e.g., Sherblom-Woodard, 2002). This online register was created to reify social stratification by differentiating the "elite" from newcomers, but it also avoided content filters imposed by (outside) moderators. The use of numbers and symbols in place of visually similar letters made leetspeak legible enough to humans to retain its communicative power while obfuscating enough content to bypass filters. One prominent example of language play online as protective self-censorship is the linguistic creativity of social media users responding to China's government-controlled media censorship. Because "Chinese netizens are still speaking in a heavily monitored environment…their demands for greater freedom of information and expression often find voice through coded language and metaphors that allow them to avoid outright censorship" (Link & Qiang, 2013, p. 79). For instance, users of online platforms like Weibo can use a romanized orthography, Korean orthography, Chinese characters with emoji interspersed, or four-digit telegraph codes that formerly corresponded to characters (Bandurski, 2020), and these coded forms often exploit the tonal system of Chinese varieties (China Digital Times, 2015).

"Community Guidelines," Content Moderation, and Algorithmic Inequity

Linguistic self-censorship by platform users is just one consequence of content moderation and algorithmic systems in digital space. A comprehensive discussion of the impacts of either is beyond the scope of this article, but critical research across media studies, human-computer interaction, and science and technology studies (among other fields) has demonstrated consequences across national, cultural, and digital contexts. This includes the reinforcement of global systems of oppression such as white supremacy, anti-Blackness, and colonial logics. As Benjamin (2019) argues, racist logics are embedded into any automated technology that does not actively resist a status quo approach, because white supremacy is the status quo: From facial recognition to recommended internet search results to spell check and autocorrect, "the raw data that robots are using to learn and make decisions about the world reflect deeply ingrained cultural prejudices and structural hierarchies" (p. 59) (see also, e.g., Sap et al., 2019). Even in the realm of manual content moderation, discriminatory ideologies that individuals are socialized to believe inform their decision-making about "appropriate content." For instance, white people are more likely than people of color to subscribe to a neoliberal "colorblind" ideology that asserts that acknowledging and discussing race necessarily perpetuates racism (Bonilla-Silva, 2018). This means that white content moderators making decisions about what constitutes "hate speech" may be more likely to deem posts with *white (people)* and related words as such. These ideological differences are also important because users who are not official content moderators still have the power to report videos that they believe violate platform guidelines (or use that as cover to report videos and/or creators they dislike). TikTok states that it "proactively enforces" guidelines to pre-empt users having to report videos, but users' beliefs about "(un)acceptable" words and ideas may differ from moderators' beliefs. Gillespie (2018) points out that community guidelines are initially written based on platform operators' assumptions about community values; and while these guidelines have the potential to be a "living document" that is revised in response to user feedback, they ultimately remain within the control of a few people. They will never represent all users' values, and there will always be contested gray areas – problems exacerbated as a platforms' user base increases and diversifies.

Recent research has highlighted the disproportionate negative impact of social media content moderation on already-marginalized populations. In their report on content moderation on Facebook, YouTube, and Twitter, Diaz and Hecht-Felella (2021, p. 3) found that policies "are drafted in a manner that leaves marginalized groups under constant threat of removal for everything from discussing current events to calling out attacks against their communities." These policies are enforced at the discretion of the platforms' executives in ways that protect the powerful and maintain the platforms' public images, often at the expense of marginalized communities. Twitter ignored Black women's evidence of racist and misogynistic trolling and was one of the last major platforms to address QAnon conspiracy theories (Diaz & Hecht-Felella, 2021). YouTube has demonetized the platforms of LGBTQ+ creators (Sung, 2019) and TikTok has algorithmically suppressed the content of queer, trans, and disabled creators instead of mitigating the harassment they experience (Rauchberg, 2022). Ohlheiser (2021) discusses how multiple platforms have been called out for suppressing marginalized creators for "false positive" guidelines violations – i.e., content that is not a violation but is interpreted as such – but TikTok's highly secretive algorithmic structure makes responding to this issue particularly difficult.

The linguistic self-censorship we analyze here is one strategy that TikTok creators from marginalized groups employ to avoid these "false positive" violations and resist what they perceive as the platform's attempt at algorithmic control of their content. In Peterson-Salahuddin's (2022) terms, linguistic self-censorship is one of the "digital tactics employed to evade detection and inverse structures of power" (p. 2) in digital space. These tactics of "digital dark sousveillance" (Peterson-Salahuddin, 2022) stem from creators' understandings of how platforms operate and how their algorithmic systems surveil the activities of racialized and otherwise marginalized subjects. Karizat et al. (2021) found that users altered who they interact with, how they interact with them (e.g., liking, commenting, or rewatching videos), and how they create content (e.g., mismatch video and audio) based on their algorithmic folk theories about this phenomenon. We expand understanding of the third category by not only identifying censored forms that creators use (e.g., Klug et al., 2023), but also by analyzing the specific linguistic processes creators leverage to render their content less algorithmically detectable.

Data and Methods

Data Collection

Because of the nature of the phenomenon and the FYP-centric structure of the TikTok platform, our data collection was limited in ways outside of our control. A typical self-censored video is designed so that it should not appear in the results returned when someone searches in the app for the censored target word (e.g., a video in which *sex* is written as *sects* should not appear in video search results for "sex"). Therefore, we were limited to collecting videos as they appeared on our FYPs. Both authors have been active TikTok users since 2020, which has allowed us to observe this phenomenon over years; however, this has also narrowed the scope of our FYPs since the algorithmic suggestions have become more tailored to our viewing habits.

Prior to this study, we established a system for collecting and coding TikToks that interested us as linguists, using the media organization app Odin (https://onodin.com/) to export and organize videos into folders directly from the TikTok app. As seen in Figure 3, self-censorship emerged as one of many discourse practices of interest.

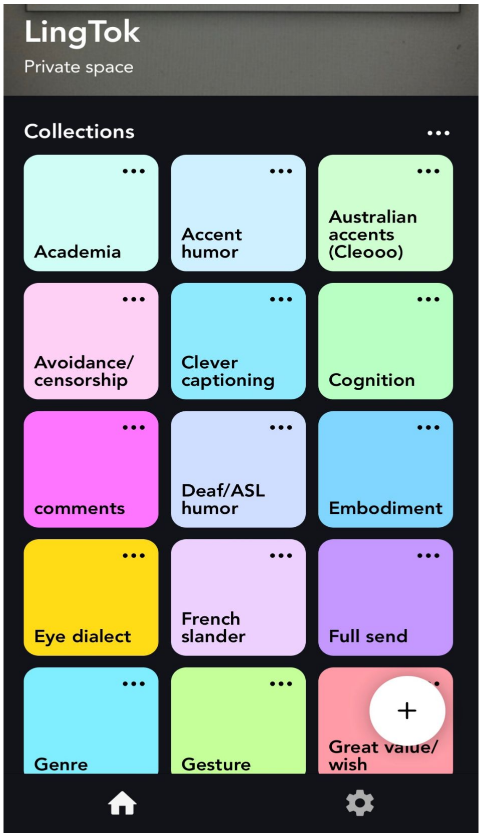


Figure 3. Folders ("collections") in Odin used to thematically organize TikToks

For the present analysis, we used Odin to compile TikToks that included self-censorship between May 2021 and February 2022, resulting in a data set of more than 200 English-language videos. (Because of how the FYP operates, our data set includes videos posted prior to this timeframe). The majority of these videos were in varieties of U.S. English, a product of the algorithmically recommended content we saw as U.S.-based English speakers. The total number of videos containing self-censorship that we encountered during this time exceeds those compiled in Odin, but many were repetitions of the same form, so we chose not to duplicate them in the dataset. Our goal was for the data to reflect the range of censorship forms rather than create a dataset representative of their relative frequency – the latter being impossible given the partial perspective that the phenomenon and platform create. We had no restrictive criteria for initial data collection: Any video by any creator that contained linguistic self-censorship of any word(s) was eligible for inclusion in the dataset.

Analytical Methods and Data Representation

As we collected examples, we noted patterns in the words that were censored, the linguistic processes used to get from the target words to the censored forms, and the identities of the creators using censored forms. Therefore, we had already identified some preliminary themes and important contextual information before we formally analyzed the data through a combination of humanistic coding and discourse analysis. We first surveyed the approximately 200 videos to broadly code for (1) topic of the censored word (e.g., race, drugs) and (2) process(es) to create the censored form (e.g., change a sound). To analyze a linguistic process, we used video context to understand what the censored word was, then compared the form in the video to the standard form of the word. We then determined what structural changes – phonological, morphological, orthographic – occurred to get from the standard to the self-censored form and whether they were already established linguistic processes.

The examples included here were chosen to illustrate the range of linguistic processes and their applications represented in our data. Because of varying permissions from users, the rapid spread of trends on TikTok, and the in-group and intertextual nature of much TikTok content, examples are not all represented or analyzed in the same ways. In some cases, the original video in which an example occurred or the creator's entire account have been deleted since our data collection; because we do not know the reason for the removal, we intentionally include the data only as an unattributed example in a table. Other examples are not analyzed simply due to space considerations. Screenshots, account handles, and creators' names included in the discussion of examples are all from public accounts with at least one million followers and/or included with the creator's consent. Where individual attribution for a popular form or trend is not possible but broad attribution to a community of users is, we include that in our discussion.

Content Warning and Researcher Positionalities

Before we turn to the analysis, we want to offer a content warning that the data we examine include multiple slurs. Because we are explaining the various linguistic processes that transform the original word into its censored form, we believe that it is important that the original word is represented in its entirety. In our discussion, we explain how these are instances of reclaimed in-group uses of the words, but that context is not readily visible in the summary tables. We also recognize that the idea of reclaiming derogatory language is contentious, since not all members of a marginalized community agree that a term can be reclaimed or who is permitted to use it (e.g., Brontsema, 2004). While we do not belong to all the groups to which the represented words have been directed, as a Black woman (Calhoun) and a white queer woman (Fawcett), we both identify as members of groups marginalized by race, gender, and/or sexuality. We understand the potential harm caused by any representation of derogatory language, especially by outgroup members, and want to approach this reality with care. Our goal is not to normalize the representation of slurs for linguistic analysis, but rather to offer the most complete picture possible of the complex linguistic reality that TikTok creators from marginalized groups must navigate on the platform.

Censored Forms and Community Guidelines

Tables 1 and 2 provide examples of words in our data with a high possibility of being flagged by automatic detection and/or human content moderators, paired with examples of self-censored forms TikTok creators have used to avoid producing those words. The target words in Table 1 have clear possibilities for use in ways that violate guidelines. The target words in Table 2 have greater potential to be "false positives" based not just on context of use, but differing ideologies about their referents. The examples in both tables are from posts in which the creator censors the target word even though it is not used in the spirit of harm or illegality that the community guidelines are intended to prevent.

For instance, *drugs* is self-censored despite the word potentially referencing legal, over-the-counter pharmaceuticals. Weed is the most frequently referenced drug, and it is self-censored despite the legalization of medicinal and recreational cannabis throughout the U.S. References to sex work(ers) are also self-censored. Although discussing sex work is not the same as showing nudity or sexual activities, self-censorship of this content reflects creators' awareness that content moderation policies can "[replicate] puritan, conservative values that conflate…sex with a lack of safety, and a lack of safety with women's bodies" (Are, 2021, p. 14). Descriptors for marginalized groups are also self-censored, even in neutral content, likely accounting for the belief among some that an outgroup member acknowledging someone's membership in a marginalized group is itself a derogatory act. In addition to homophobic beliefs about queer sexuality as deviant, stereotypes of queer people as hypersexual could also contribute to self-censorship of words like *gay* and *lesbian,* since using them could be construed as promoting or normalizing "inappropriate" sexual behavior (e.g., Nadal et al., 2016).

Self-censoring certain words can also be a tongue-in-cheek way to bring attention to ideologies. The word *nipples*, for example, may be erroneously flagged via automatic detection based on the same logic about women's bodies that would deem *sex worker* inappropriate. A creator's choice to self-censor could also be interpreted as them highlighting the over-policing and hypersexualization of women's bodies. For example, in one video, a creator uses TikTok's greenscreen function to display an image of a woman modeling a dress; as they describe the clothing, the creator suggests the image is misleading because "they edited out her nip nops," unnecessarily self-censoring *nipples* to *nip nops.*

Community guideline category[2]	**Target word(s)**	**Censored form**
Hateful speech and ideologies	*nigga(s)*	*knick knack patty whack(s)* *ninja(s)*
	Nazi	*n@z1, not-see*
Suicide, self-harm, and violent acts	*kill*	*unalive*
	suicidal	*sewer cidal*
Illegal activities and regulated goods	*drugs*	*droogs*
	weed	*ouid,* 🍃
	cocaine	❄️
Adult nudity and sexual activities	*sex worker*	*accountant*
	BDSM	*BeatySM*

Table 1. Words likely to be flagged by automated and/or human content moderators

Community guideline category	**Target word(s)**	**Censored form**
Hate speech and ideologies	*white*	⚪, 🦷
	white supremacy	*YT soup remassy*
	homophobic, homophobia	*hydrophobic, cornucopia*
Harassment and bullying; Adult sexual activities	*gay*	*gäÿ, ghey*
	lesbian	*le$bean*
Adult nudity	*nipples*	*nip nops*

Table 2. Words that may be erroneously deemed violations based on differing ideologies

Now that we have illustrated the relationship between TikTok's community guidelines and creators' self-censorship practices (the "why" of this phenomenon), we turn to an analysis of the "how": the various linguistic processes used to create these self-censored forms.

Analytical Findings

TikTok creators' linguistic processes for self-censorship mirror the manipulation of sound, spelling, morphology, and meaning used in documented forms of language play and self-censorship in offline and online contexts, as discussed above. However, there are crucial differences based on goals, contexts, and language structures. For example, like social media users in China, TikTok creators in our data play with sound to avoid using specific linguistic forms, but the former may manipulate tone whereas the latter may play with syllable patterns. Unlike predictable forms of language play like pig Latin, there are no consistent rules for TikTok self-censorship; some forms become widely popular, but each user chooses whichever strategies they prefer.

Key to TikTok self-censorship is maintaining legibility of meaning for the right audience(s) regardless of the strategies used to change the linguistic form. Videos should circumvent automated language detection, but in-group TikTok users who see the videos on their FYPs should be able to understand what is being communicated. Therefore, we analyze TikTok creators' strategies for self-censorship as a repertoire of linguistic resources whose communicative effect is akin to that of avoidance registers (e.g., Haviland, 1979). We describe and analyze seven categories of linguistic processes for text-based forms represented in our data and then discuss strategic uses of gesture and voice quality that co-occur with some of these text-based processes.

Linguistic Processes to Create Censored Forms

We identified seven categories of linguistic processes in our data, represented in Tables 3a and 3b. These categories reflect attested tendencies in language play, such as the inclusion of onomatopoeia and acronyms (Sherzer & Webster, 2015), as well as new practices specific to technologically mediated communication. Some categories include only one process, and other categories encompass multiple processes. We include more examples in the tables than we discuss in order to illustrate the varied ways that different creators engage in the same linguistic process. Additionally, some target words appear as examples for multiple processes (e.g., *gay, white*) to demonstrate that creators may choose different self-censorship strategies for the same target word. The linguistic processes we have identified are not exhaustive of all possible processes creators employ, but rather demonstrate the range of linguistic resources and creativity involved in this practice.

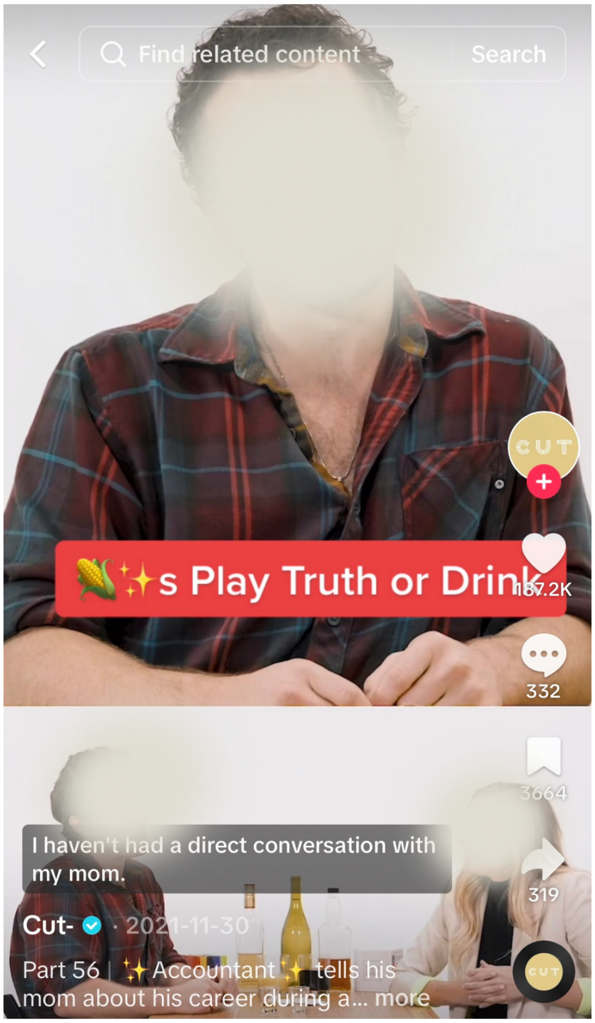
	Process	**Explanation / Description**	**Examples**
1.	Use of non-letters	Numbers, symbols, diacritics, spaces, emoji	*n@z1* ‘Nazi’ ⚪,🦷 ‘white’ 👉👌 ‘sex(ual)’ 💅, *gäÿ* ‘gay’ *le$bean* ‘lesbian’ sm o ke cr a ck ‘smoke crack’
2.	Innovative use of English morphology	Using existing morpheme in novel way	*unalive* ‘kill’
3.	Lexical replacement	Semantic dissimilarity	*accounting* ‘sex work’
		Semantic dissimilarity with phonetic similarity	*baguette* ‘faggot’ *shrek work* ‘sex work’
		Semantic similarity	*colorless, palm colored* ‘white’ *8.5 x 11s, blank Google docs* ‘white people’
4.	Innovative phonological patterns	Applying attested phonological rule to unlikely word	*s*🥚🥚, i.e., *s-eggs, ‘sex’* *seggsy, sessy* ‘sexy’
		Swap vowels within phonemic inventory	*droogs* ‘drugs’ *fuhgoot* [fəˈgut] ‘faggot’

Table 3a. Linguistic processes to create censored forms

The first category of self-censorship processes, the use of non-letters, includes the use of numbers, symbols, spaces, and diacritics – available since the plaintext era of the early internet (McCulloch, 2019) – as well as the more recently invented emoji. For example, *n@z1* uses one-for-one substitution of the symbol ‘@’ and number ‘1’ for the letters ‘a’ and ‘i,’ respectively, to avoid spelling *Nazi*. The emoji ‘💅,’ in contrast, replaces the entire word *gay*, working as an iconic (i.e., visually similar) representation of an indexical embodied practice (Gal & Irvine, 2019; Silverstein, 2003) based on Western stereotypes of effeminate male homosexuality.

Process two, innovatively using English morphology, includes using the existing English prefix *un-* in novel ways. The word *unalive*, meaning ‘kill,’ features the reversative form of the prefix (Zimmer et al., 2011), but unlike its typical uses that do not change the word’s lexical category, here it derives a verb from an adjective. Notably, the resulting innovative form is productive in ways that follow English morphological rules, e.g., “When your therapist asks if you have thoughts of unaliving yourself.”

The third category, lexical replacement, includes three different processes: replacing a word with another that (1) has a completely unrelated meaning, (2) sounds similar but has an unrelated meaning, or (3) has a related meaning. One widely circulated example of lexical replacement with semantic dissimilarity is *accounting/accountant* to refer to sex work(er) (Figure 4): Other than being another possible occupation, accounting has nothing obvious in common with sex work.



I haven't had a direct conversation with my mom.

Cut- · 2021-11-30

Part 56 | ✨Accountant✨ tells his mom about his career during a... more

Figure 4. Video from @cut with the caption: “✨Accountant✨ tells his mom about his career during a game of #TruthOrDrink”

Stereotypes of accounting as a boring, steady job make it an ideal cover for people who do not want to talk about their actual jobs. In July 2020, creator @rockysroad posted a musical video about telling people they are an accountant to avoid questions about their acting career; sex workers, among others, used the song as a sound in their own TikToks, visually hinting at their real jobs. A topically related example of semantic dissimilarity with phonetic similarity is *shrek work* in place of *sex work*. This process is similar to Cockney rhyming slang, a well-documented form of sound-based language play that "hides" a phrase using unrelated rhyming (or near-rhyming) words (Green, 2003).

The third form of lexical replacement uses words with semantic similarity regardless of phonetic similarity. Like the creation of many new slang terms, this process involves using existing English words – as they are or combined into novel expressions – for new referents. Also, like slang, distinct forms of self-censorship can develop among different communities of TikTok users. Black TikTok creators popularized creative words for white people, a practice adopted by creators from other racialized groups. These self-censored forms rely on the shared quality of "white(ness)," though the exact type of whiteness is not the same. These forms collapse "white" the color and "white" the racial category, allowing white-colored referents to be linguistic stand-ins for whiteness and/or white people. As iconic linguistic signs (see Gal & Irvine, 2019) for white people, *8.5 x 11s* and *blank Google docs* use the whiteness of blank paper as the basis of similarity between the phrase and the referent, e.g., "My dog came from a kennel run by 8.5 x 11s." We also see this semantic relationship in the use of emojis of white-colored objects to replace the word *white.* Responding to another user's comment – "can you stop fucking calling us white people" – one Black creator edited their automated speech transcription to read, "I was acknowledging that this person in that story was 🧻." Some creators of color also describe white people as "people without color" or "people of no color" to parallel "people of color" in a way that decenters whiteness. *Without color* can be stated more concisely as *colorless,* leading to *colorless* as a censored form for *white*.

Category four, innovative phonological patterns, also includes multiple processes. The first set of examples involves innovative application of existing phonological rules to unlikely words. In these cases, documented sound patterns in English are used in contexts where they do not typically occur but are not phonologically restricted. For example, *seggsy* [sɛgzi] includes the recognizable word *eggs* to indicate a pronunciation of *sexy* that applies intervocalic voicing, a form of phonological assimilation, to the voiceless consonant cluster [ks]. Although intervocalic /s/-voicing of certain words is an attested dialect feature in U.S. English (e.g., *greasy* [gɹizi]; Wolfram & Schilling, 2015), the changed spelling suggests that this is intended as an unusual pronunciation.

The next set of examples creates innovative sound patterns by swapping the usual vowel(s) in a word with other phoneme(s). These self-censored forms are therefore plausible English words but often play with the sound-spelling relationship or typical stress patterns. For example, the censored form *droogs* [dɹugz] for *drugs* replaces 'u' /ʌ/ with 'oo' /u/, drawing on shared sound features (backness) and the orthographic variability for these sounds in English in its different spelling and pronunciation.

	Process	Explanation / Description	Examples
5.	Intentional spoonerism	Metathesis of word-initial sound(s) in a phrase	*woke smeed* 'smoke weed'
6.	Orthographic reanalysis	Word "respelled" with real homophones	*sir-come-sized* 'circumcised' *knee grow* 'negro' *sects* 'sex' *not-see 'Nazi'*
		Words respelled with fake homophones (nonce words following English phonology)	*YT soup remassy* 'White supremacy' *sewer cidal* 'suicidal' *beatySM* 'BDSM' *ghey* 'gay' *raycest* 'racist' *keenk* 'kink'
		Initialism converted into real word(s) or phrase(s)	*leg booty* 'LGBT'
7.	Phonotactic/ prosodic template	Replace target with form of equal syllables and stress placement as well as shared segments	*hookedonphonics* 'homophobic' *hydrophobic* 'homophobic' *homophonic* 'homophobic' *cornucopia* 'homophobia'
		Ablaut reduplication	*nip nops* 'nipples' *nig nogs* 'niggas'
		Rhyme reduplication with assimilation	*mugga chugga* 'mother fucker' [mʌðə fʌkə]

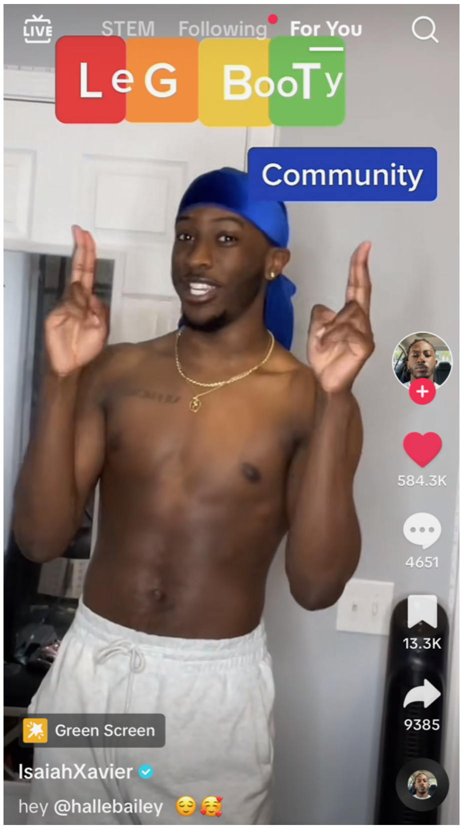
Table 3b. Linguistic processes to create censored forms, continued

Examples in category five, spoonerisms, draw on patterned errors that reflect the structural properties of spoken language: Speakers swap initial sounds of two or more words in a phrase but retain the material following those sounds. The unintentional process that leads to these forms in natural spoken language is employed on TikTok intentionally to create nonce, but analyzable, forms. For example, in one video Demetrius Fields (@demetriusfields) metathesizes the initial sounds [sm] and [w] in the phrase *smokin so much weed* to create *wokin so much smeed.* This case differs from most unintentional spoonerisms since the resulting forms are not all known or attested words; however, it still relies on linguistic knowledge of not only the phenomenon but also of syllable structure and phonotactics. He swaps the onset [sm] as opposed to solely the initial [s], which conforms to the syllable structure and avoids an unattested initial sequence (in English) of [wm].

Category six, orthographic reanalysis, is based on intentional misspellings and/or mispronunciations of words with some phonetic similarity to the original form. The first two types of orthographic reanalysis involve “respelling” words with homophones in ways that parallel the linguistic process of the popular game Mad Gab (Mattel, 2020) and the phenomenon of “eggcorns.” Mad Gab creates phrases out of phonetically similar but orthographically different words; players have to read the homophonous phrase in order to hear the target phrase, as one can read *sir-come-sized* and hear *circumcised*. Eggcorns have the opposite directionality: They occur when a listener mishears speech or reinterprets something as a homophone and writes it as such, like *egg corns* for *acorns* (Liberman, 2003) or *beatySM* for *BDSM*. Like intentional spoonerisms, this type of reanalysis turns a naturally-occuring language error into a linguistic choice.

For self-censorship purposes, the homophones need not be real words as long as the orthographic form can elicit a pronunciation that approximates part of the target phrase. Reanalysis can flout structural properties of the language by changing stress or dividing consonant clusters by shifting syllable or word boundaries. One creator uses this strategy when they write “values of YT soup remassy,” where *soup* is the only part one would find in a standard dictionary. In the first component of the censored form, ‘y’ [waɪ] is read as the name for the letter and ‘t’ [t] as the sound the letter tends to make; together they are pronounced [waɪt] ‘white.’ The second part of the target, *supremacy,* has more than one accepted pronunciation [səˈprɛməsi, suˈprɛməsi]; based on the latter, *soup* [sup] works as a partial stand-in, crossing syllable boundaries and leaving [rɛməsi] to be represented as *remassy.* Although ‘c’ can be realized as [s] or [k], it consistently surfaces as [s] before ‘y,’ meaning this orthographic change from ‘c’ to ‘ss’ was not necessary in this nonce form; it may have been done for further obfuscation or to index the creator understands this process as a show of creativity.

The third type of orthographic reanalysis concerns initialisms where the phonetic form is embellished. When queer creator Isaiah Xavier (@isaiahxavier10) describes himself as “a member of the leg booty community,” he plays off the initialism LGBT (read as letters) and adds vowels to create a readable English word (Figure 5). This process is not turning an initialism into an acronym (read as a word), but instead adds segments to the initialism to create a word that is no longer an abbreviation. *Legbooty* has been widely circulated enough to become its own lexicalized form that functions as a one-for-one substitution for *LGBT* and longer forms of the initialism (Wright et al., 2023).



IsaiahXavier

hey @hallebailey 😌🥰

Figure 5. “LeG BooTy community”

Finally, category seven is the use of a phonotactic or prosodic template. In the first subtype, creators replace the target with a form that has a comparable number of syllables but can also share stress placement and sounds. For example, *hookedonphonics* [ˌhʊkt.ɑnˈfɑ.nɪks] as a self-censored form for *homophobic* [ˌhoʊ.məˈfoʊ.bɪk] conserves the number of syllables and primary and secondary stress placement (σ̀σσ́σ); it also includes the shared sounds [h] and [f]. The template – ˌhV(CC).(C)V(C).ˈfV.CVk(C) – shows how other avoidant forms (e.g., *hydrophobic*) also conform.

The remaining templatic subtypes rely on a strategy of complete or partial reduplication and some form of antiphony – repeated syllables that differ by a single sound. The first, *nip nops* ‘nipples,’ is an example of ablaut reduplication like *riff-raff* or *zigzag*. These forms follow a prosodic template in which the first syllable carries primary stress and features a high front vowel, while the second syllable has a low vowel. While perhaps infrequent today, this has been a marginally productive strategy for English word formation, especially for words with mildly pejorative meanings (e.g., Minkova, 2002). The second type of reduplication, represented by *mugga chugga* [mʌgə ʧʌgə] for ‘mother fucker,’ is reminiscent of rhyme reduplication (e.g., *hocus pocus*), but there are multiple influences on the form, which was produced by a Black American creator. First, the final vowel is non-rhotic, a distinguishing feature of U.S. Black English varieties, in which the target word *mother fucker* would frequently be pronounced [mʌðə fʌkə]. Second, the use of ‘gg’ [g] instead of the target sounds ‘th’ [ð] or ‘ck’ [k] is a form of consonant harmony – namely, velar assimilation where apical consonants assimilate to nearby velar consonants, as is common in child language (e.g., *tickle* [gigu]; Ingram, 1986). Finally, the initial ‘f’ of *fucker* is changed to ‘ch,’ perhaps in analogy with *chugga chugga* [ʧʌgə ʧʌgə], the widely recognized imitative form for a moving train in English. Thus, in this one form we see the use of a variety-dependent vowel pattern while also appealing to widely used patterns in English.

Additional Categories of Censored Forms

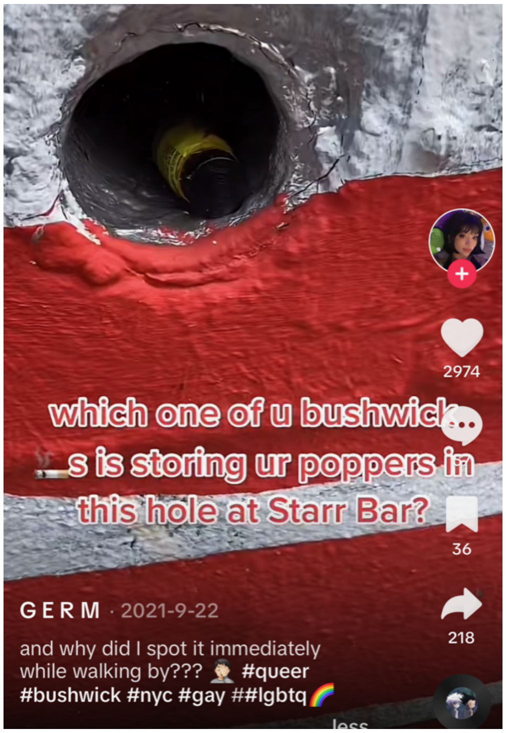
These seven categories of linguistic processes represent a sample of strategies that TikTok creators use to change a word’s form while maintaining legibility of meaning in context. In each of the examples discussed above, the transformation from target word to censored form occurs in a single step. In other cases, the transformation results from successive steps, i.e., applying additional processes to an already censored form that continues to circulate. For example, the popular form *le-dollar-bean*/*le dollar bean* resulted from the combined processes of substituting non-letters, reanalyzing spelling to create a fake homophone, and the orthographic representation of text-to-speech pronunciation. The initial censored form, *le$bean,* uses the common substitution of $ for ‘s’ and changes the spelling for the last two syllables from *bian* to *bean*. This change depends on a reanalysis of *bean* as two morphemes, *be* and *an* [biɪn], rather than the single morpheme *bean* [bin]. When one creator applied the text-to-speech technology to *le$bean,* the word was parsed as le-$-bean [lɛ dɑ.lɚ bin]. This pronunciation was then represented orthographically as *le-dollar-bean,* which has become lexicalized and now co-exists with *le$bean* as a censored form for *lesbian.*

Another category of censored forms are ones that are changed in a single step but require intertextual or interlingual knowledge to interpret, such as knowledge of established internet discourse practices or other language varieties, as demonstrated by the examples in Table 4.

Target word	**Intertextual/Interlingual knowledge**	**Censored form**
vibrator	🍆 ‘penis’ *spicy* ‘sex-related’ or ‘non-normative’	*Spicy eggplant*
fag(got)s	*fag* ‘cigarette’ in British English	🚬*s*

Table 4. Censored forms requiring intertextual or interlingual knowledge

To arrive at 🚬*s* for *faggots,* the target is first shortened to *fags*, which creates the possibility of polysemy for English speakers who know that the word refers to cigarettes in varieties of British English. This alternative definition is represented by the cigarette emoji for singular *fag*, which is pluralized by -s. For example, Gem (@glitterboii) asks in a video, “which one of u bushwick 🚬s is storing ur poppers in this hole at Starr Bar?” and uses hashtags (#queer, #gay, #lgbtq🌈), which make clear this is an in-group use of the reclaimed meaning by a queer person and not a derogatory outgroup use of the slur (Figure 6).



which one of u bushwick 🚬s is storing ur poppers in this hole at Starr Bar?

GERM · 2021-9-22

and why did I spot it immediately while walking by??? 🤦 #queer #bushwick #nyc #gay ##lgbtq🌈

Figure 6. Use of emoji 🚬 for *fag*

The way this form exploits the polysemy of *fag* across varieties of English is comparable to “interlingual puns” (Sherzer & Webster, 2015). These are puns “based on the phonological iconicity of forms that cross linguistic boundaries” (p. 4), i.e., forms that are shared by multiple linguistic systems and correspond to different meanings. Another censored form that could be considered an interlingual pun is *ouid* for *weed*, which relies on the similar pronunciation of the French *oui* and the English *wee*.

Multimodal Self-Censorship

Our analysis thus far has focused on speech and text forms, but TikTok’s audio-visual features afford multimodal creativity. Some creators self-censor through gesture instead of, or in addition to, speech and text; however, gestural strategies appear to be overall less frequent, despite embodied performance being foundational to TikTok since its beginnings as the lip-synching app Musical.ly (e.g., Rettberg, 2017). Using gesture as a censorship strategy seems to require more context and/or conventionalized use (see *palm colored*), making it more difficult to generate legible novel gestural forms. Gestural strategies vary in terms of how much they actually “censor” a target form, but they all function to signal creators’ social positionings vis-a-vis ideologies about “controversial” topics. We briefly turn to two examples to demonstrate the highly contextual nature of self-censored forms that successfully leverage gesture.

In one video, a trans woman makes use of non-modal voice and manual gesture. When she says, “I didn’t have the heart to tell her that I don’t have periods because I’m trans,” she whispers *trans* and simultaneously blocks her mouth with her hand. These are existing communicative practices in offline discourse: whispering content you do not want others to hear and covering your mouth to prevent someone from reading your lips. The speech stream remains audible, and *trans* appears in the video's hashtags, so the word is not actually "censored." Rather, the creator's performance of self-censorship indirectly points to the contentious nature of trans identity in U.S. society.

The gesture of holding one's palm to the camera as a stand-in for *white* involves a semiotic strategy similar to forms like *8.5 x 11s* as a lexical replacement for *white people*. This now-conventionalized gesture is the embodied equivalent of the term *palm colored* and the palm emoji (✋🏻), two forms that appear in text. Without the textual referents, the palm gesture could potentially be ambiguous since people of all skin tones have light palms. However, Black creators, who popularized all three forms, developed a way to make clear who the gesture refers to: In multiple videos comparing things Black and white people do, Black creators point to the back of the hand or wrist to indicate Black people/Blackness and the palm when referring to white people and/or whiteness.

Language Play for Cachet

A final important finding in our data was the phenomenon of self-censorship as a form of "language play for cachet": TikTok creators censoring words for the primary purpose of performing their linguistic skill, participating in a trend, and/or indexing sociopolitical awareness. Given the value of linguistic creativity on TikTok and the platform's memetic culture, it is not surprising that this form of self-censorship eventually developed. The censored forms analyzed above result from creators understanding the different ways that words with ideologically fraught referents may be linked to TikTok's community guidelines. Self-censorship for cachet, however, targets words that have no legitimate ideological grounds for censorship. One word that reflects this phenomenon is *Italian* and its various censored forms (e.g., *Ital!@n*), which reference a recurring topic of discussion online: that recently some white Italians have tried to claim status as people of color to position themselves as racially, and therefore socially, marginalized in the U.S. In contemporary U.S. society, racial ideologies and categorizations are such that *Italian* – when referring to ethnic/cultural background rather than nationality – means "white" unless specified otherwise. Following strategies noted above, self-censoring *Italian* indirectly references these discourses by satirically treating the word as a contentious ethnoracial label like *white.* This phenomenon is also reflected in *Italianx*, a facetious analogy to *Chicanx* and *Latinx* without the goal of gender-neutral morphology. "Censoring" *Italian*, then, is a stance-taking move: Creators who do so evaluate the belief that people of Italian descent are necessarily people of color as laughable, distancing themselves from the belief and the people who espouse it (cf. Dynel [2020] on "disparaging humor").

Discussion

Linguistic self-censorship on TikTok reflects the tensions between the legitimate need for content moderation online and the social biases built into technology designed to do the moderation. Beyond a perceived necessity to avoid being "guidelined," a creators' choice to self-censor on the platform can be seen as an indirect call out (see Clark, 2020) of TikTok's unequal enforcement of guidelines, as well as the ideologies (around race, gender, sexuality, etc.) that make certain topics or words potential violations of those guidelines. That creators from marginalized groups disproportionately self-censor contradicts TikTok's assertion that their "Community Guidelines establish a set of norms and common code of conduct that provide for a safe and welcoming space *for everyone*" and that they "prioritize safety, diversity, inclusion, and authenticity" (TikTok, 2022, emphasis added).

TikTok is not the first platform to fall short of its stated promises to protect users from harmful behaviors and enforce its community guidelines fairly: Facebook, YouTube, and Twitter have also been found at fault (Diaz & Hecht-Felella, 2021). TikTok's status as one of the most influential social media platforms globally, however, has generated greater attention to its content moderation policies, users' algorithmic folk theories, and creators' discursive choices (self-censorship and otherwise). Additionally, TikTok is the first major video-based platform where linguistic self-censorship has occurred in the public eye in text, speech, and gesture – compared to primarily text-based practices on preceding platforms (e.g., leetspeak) – and one of the first platforms to have mimetic discourse practices structurally encouraged (Zulli & Zulli, 2020). These factors have converged to make linguistic self-censorship on TikTok a particularly salient iteration of a decades-long phenomenon in online discourse.

As a form of language play, self-censorship on TikTok draws on linguistic practices that predate the internet: manipulating sounds, spelling, syntax, and meaning to create new linguistic forms for the sake of fun and/or to avoid consequences based on top-down enforcement of hegemonic ideologies. Processes such as ablaut reduplication and prosodic templates require knowledge of English phonology. Lexical replacement and orthographic reanalysis are possible through knowledge of English's lexicon and sound system, as well as common words in other languages (e.g., *ouid* for *weed*). Like its predecessors in other digital communities, self-censorship as language play on TikTok relies on users' ability to apply this knowledge to the unique combination of features the platform offers. Through audio, video, image, and text, along with features like text-to-speech, TikTok offers creators an expansive toolkit to create self-censored forms, including speech, writing, emoji, and gesture. Because of the many possibilities that this generates, there is no formal set of rules to create new self-censored forms on TikTok – contrasting platform vernaculars with identifiable, if fairly open-ended, patterns like the morphosyntactically playful "doge" register (see Brook & Blamire, 2023). A single form can also continue to morph over time by undergoing different types of transformations (e.g., *lesbian* to *le dollar bean*). TikTok's memetic culture and the general value of linguistic creativity on the platform has led to the expansion of self-censorship from necessary, based on real or perceived threat of content moderation, to fully playful, as participation in a platform trend and a means to index sociopolitical alignments. Rather than the latter replacing the former, these functions now co-exist on TikTok, creating a continuum of playfulness.

This means that different creators may use the same linguistic form with different intents, but their respective intents may not be discernible to viewers. As researchers we have aimed to carefully interpret examples of linguistic self-censorship based on linguistic context, sociocultural and political context, and our familiarity with TikTok culture and self-censorship practices from our experiences as users; however, we are not immune to misinterpretation. In our correspondence with one TikTok creator, they explained that our interpretation of their use of the term *los jibbities* as a self-censored form of the initialism *LGBT* was inaccurate. They were referencing a series of videos by @dez.thelez joking about her mother's creative language use that often involves her Spanish language knowledge (in this instance, *LGBT > el GBT > los GBTs > los jibbities*). Given similar censored forms in our data – interlingual and initialism-based forms, censored forms of *LGBT* and related terms (*gay, lesbian, trans*) – and a censored form (*b*tches*) in the caption of the creator's video, *los jibbities* was a highly plausible self-censored form in this context. This type of misattribution of intent is different from misinterpreting words because they are ineffective self-censored forms. Similar to the continuum of playfulness, there is a continuum for effectiveness: Whereas some forms are immediately identifiable as censored (*b*tches*), others are legible with sufficient context and/or in-group knowledge (*accountant*), and some are widely confusing because the connection between the target and censored forms is too niche or broad, or the censored form is too easily interpreted literally (e.g., *mascara* 'sexual assault'; Cheung, 2023). Effective self-censored forms reflect a shared understanding of what might be seen as contentious and a balance of transparency and obfuscation realized via linguistic choice and sufficient contextualization. As discussed above, there are no set rules that ineffective forms flout, but these forms fail to fully achieve this necessary balance in the context in which they are operating.

Creators produce self-censored linguistic forms in response to conditions on TikTok, but these forms travel to other platforms, to digitally-mediated communication like texting, and even into offline discourse – meaning a single TikTok video or trend can have a wide reach beyond the millions of potential viewers on the platform alone. TikTok's existence within a social media ecosystem means that its content is influenced by content on other platforms, and the origins of certain forms may be blurred. For these reasons, we have not made claims about whether specific forms originate on TikTok; instead, we have focused on how and why they are created in the TikTok context.

The examples we have analyzed are tied to ideologies about topics long-considered controversial. There are, however, self-censored forms that are based on rapidly changing events or moments in time. This is exemplified by censored forms related to the COVID-19 pandemic. In the U.S., COVID misinformation morphed into conspiracy theories, and adherence to masking, social distancing, and vaccine mandates became politicized, creating a new set of ideologically contentious words. The most vehement vaccine and mask opponents frame government mandates as oppression infringing on individual freedoms, and themselves as targeted dissenters whose activities are unfairly monitored (e.g., Pascual-Ferrá et al., 2021). TikTok creators parodied these rhetorics by self-censoring COVID-related words. For example, in a September 2021 TikTok, comedian Caleb Hearon performed an anti-vaccine persona, saying "I wanna talk a little bit today about the c-o-v-i-d-v-a-c-c-i-n-e. Who knows what they are censoring at this point." Hearon later whispers *vaccine* and also censors the word in the caption (*v*ccine*). TikTok creators apply self-censorship practices to novel contexts like COVID in ways that illustrate an understanding of self-censorship as a register or genre with shared qualities (that are legible even in highly specific contexts) and myriad possibilities for innovation.

Conclusion

In this article, we have demonstrated how technology, ideology, and linguistic creativity have converged to create an evolving “avoidance register” of self-censored forms on the video-sharing platform TikTok. In response to unequal enforcement of community guidelines and awareness of discriminatory social ideologies, creators use the technologies of the TikTok platform to avoid algorithmic detection and censorship of their content by manipulating writing, speech, and gesture. The sample of linguistic processes analyzed in this article illustrate language’s systematicity, creators’ shared linguistic knowledge, and the range of possibilities for linguistic creativity, as well as the sociopolitical, temporal, and linguistic contexts that shape why and how creators self-censor. Linguistic self-censorship on TikTok shares qualities with language play in other contexts and self-censorship on other platforms, and our analysis contributes to a broader understanding of the complex relationship between language, sociocultural forces, and digital technology.

As one of the most influential social media platforms today, TikTok is an important site of linguistic analysis in its own right. The multimodal nature of linguistic self-censorship on the platform necessitates an understanding of language that encompasses text, speech, and gesture. English, of course, is not the only language in which self-censorship practices occur, even when limiting scope to the U.S. Studies of the processes used to create new forms in other languages, especially those whose linguistic structure and writing systems are significantly different from English, will illuminate additional ways that language and technology influence each other.

Notes

1. Much popular discussion of this phenomenon has used the term *algospeak,* based on an April 2022 article published in *The Washington Post* by reporter Taylor Lorenz. Our study predates Lorenz’s coinage of the term (our work is mentioned in the piece and many examples linked in it are drawn from our publicly available materials), and we find *linguistic self-censorship* to be a more immediately legible descriptor of the phenomenon that also contextualizes it vis-à-vis related linguistic practices, as we discuss here.
2. The community guideline descriptions in Tables 1 and 2 are those that appeared on the website during our period of data collection. TikTok updated the guidelines in March 2022.

References

Abidin, C. (2021). Mapping internet celebrity on TikTok: Exploring attention economies and visibility labours. *Cultural Science Journal*, *12*(1), 77–103. http://doi.org/10.5334/csci.140

Are, C. (2021). The shadowban cycle: An autoethnography of pole dancing, nudity, and censorship on Instagram. *Feminist Media Studies*, *22*(8), 2002-2019.

Baker, P. (2002). Construction of gay identity via Polari in the Julian and Sandy radio sketches. *Lesbian and Gay Review, 3*(3), 75–83. https://doi.org/10.1179/nam.1977.25.3.124

Bandurski, D. (2020, April 28). Skirting Chinese censorship with emoticons and telegraph codes. *Brookings*. https://www.brookings.edu/techstream/skirting-chinese-censorship-with-emoticons-and-telegraph-codes/

Basso, K. H. (1979). *Portraits of 'the Whiteman' linguistic play and cultural symbols among the Western Apache*. Cambridge University Press.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code.* Polity.

Bhandari, A., & Bimo, S. (2020). TikTok and the "algorithmized self": A new model of online interaction. *AoIR Selected Papers of Internet Research*. https://doi.org/10.5210/spir.v2020i0.11172

Biddle, S., Ribeiro, P. V., & Dias, T. (2020, March 16). Invisible censorship: TikTok told moderators to suppress posts by "ugly" people and the poor to attract new users. *The Intercept*. https://theintercept.com/2020/03/16/tiktok-app-moderators-users-discrimination/

Bonilla-Silva, E. (2018). *Racism without racists: Color-blind racism and the persistence of racial inequality in America*, 5th ed. Rowman & Littlefield Publishers.

Briggs, C. L., & Bauman, R. (1992). Genre, intertextuality, and social power. *Journal of Linguistic Anthropology*, *2*(2), 131-172

Brontsema, R. (2004) A queer revolution: Reconceptualizing the debate over linguistic reclamation. *Colorado Research in Linguistics*, *17*(1). https://doi.org/10.25810/dky3-zq57

Brook, M., & Blamire, E. (2023). Language play is language variation: Quantitative evidence and what it implies about language change. *Language, 99*(3), 491-530. https://doi.org/10.1353/lan.2023.a907010

Bucholtz, M. (2019). The public life of white affects. *Journal of Sociolinguistics*, *23*(5), 485–504. https://doi.org/10.1111/josl.12392

Calhoun, K. (2019). Vine racial comedy as anti-hegemonic humor: Linguistic performance and generic innovation. *Journal of Linguistic Anthropology*, *29*(1), 27–49. https://doi.org/10.1111/jola.12206

Cheung, K. (2023, January 27). Julia Fox didn't realize 'mascara' was code for sexual assault on TikTok. Neither did I! *Jezebel*. https://jezebel.com/julia-fox-didn-t-realize-mascara-was-code-for-sexual-1850042980

China Digital Times. (2015). Decoding the Chinese internet: A glossary of political slang (Revised edition). https://chinadigitaltimes.net/downloads/decoding-the-chinese-internet-a-glossary-of-political-slang-2015-edition/

Clark, M. D. (2020). DRAG THEM: A brief etymology of so-called "cancel culture." *Communication and the Public*, *5*(3-4), 88-92. https://doi.org/10.1177/2057047320961562

DeVito, M. A., Gergle, A., & Birnholtz, J. (2021). "Algorithms ruin everything": #RIPTwitter, folk theories, and resistance to algorithmic change in social media. *Proceedings of the 2017 Conference on Human Factors in Computing Systems*, pp. 3163-3174.

Diaz, A., & Hecht-Felella, L. (2021). *Double standards in social media content moderation.*

Brennan Center for Justice at NYU Law. https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation

Dynel, M. (2020). Vigilante disparaging humour at r/IncelTears: Humour as critique of incel ideology. *Language and Communication, 74*, 1–14. https://doi.org/10.1016/j.langcom.2020.05.001

Eckert, P. (2006). Communities of practice. In K. Brown (Ed.), *Encyclopedia of language and linguistics*, 2nd ed. (pp. 683–685). Elsevier. http://dx.doi.org/10.1016/B0-08-044854-2/01276-1

Gal, S., & Irvine, J. T. (2019). *Signs of difference: Language and ideology in social life*. Cambridge University Press.

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, *20*(12), 4492–4511. https://doi.org/10.1177/1461444818776611

Gibbs, M., Meese, J., Arnold, M., Nansen, B., & Carter, M. (2015). #Funeral and Instagram: Death, social media, and platform vernacular. *Information, Communication & Society*, *18*(3), 255-268

Gillespie, T. (2010). The politics of 'platforms.' *New Media & Society*, *12*(3), 347-364.

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press

Green, J. (2003). Rhyming slang. *Critical Quarterly, 45*(1-2), 220–226.

Halliday, M. A. K. (1976). Anti-languages. *American Anthropologist*, *78*(3), 570–584.

Haviland, J. (1979). Guugu Yimidhirr brother-in-law language. *Language in Society*, *8*(3), 365–393.

Ingram, D. (1986). Phonological development: Production. In P. Fletcher & M. Garman (Eds.), *Language acquisition,* 2nd ed. (pp. 71–92). Cambridge University Press.

Karizat, N., Delmonaco, D., Eslami, M., & Andalibi, N. (2021). Algorithmic folk theories and identity: How tiktok users co-produce knowledge of identity and engage in algorithmic resistance. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–44. https://doi.org/10.1145/3476046

Klug, D., Steen, E., & Yurechko, K. (2023). How algorithm awareness impacts algospeak use on TikTok. *Companion Proceedings of the ACM Web Conference* (April 2023), pp. 234–237 https://doi.org/10.1145/3543873.3587355

Lefkowitz, N. J. (1989). Verlan: Talking backwards in French. *French Review, 63*(2), 312–322.

Liberman, M. (2003, September 23). Egg corns: Folk etymology, malapropism, mondegreen, ???. *Language Log*. http://itre.cis.upenn.edu/~myl/languagelog/archives/000018.html

Link, P., & Qiang, X. (2013). China at the tipping point? From "fart people" to citizens. *Journal of Democracy*, *24*(1), 79–85.

Mattel. (2020). *Mad Gab*. http://prod.mattelgames.com/en-us/party/mad-gab

McCulloch, G. (2018, October 9). Welcome to Voldemorting, the ultimate SEO dis. *WIRED.* https://www.wired.com/story/voldemorting-ultimate-seo-diss-resident-linguist/

McCulloch, G. (2019). *Because internet: Understanding the new rules of language*. Riverhead Books.

Minkova, D. (2002). Ablaut reduplication in English: The criss-crossing of prosody and verbal art. *English Language and Linguistics*, *6*(1), 133–169.

Nadal, K. L., Whitman, C. N., Davis, L. S., Erazo, T., & Davidoff, K. C. (2016). Microaggressions toward lesbian, gay, bisexual, transgender, queer, and genderqueer people: A review of the literature. *The Journal of Sex Research*, *53*(4–5), 488–508. https://doi.org/10.1080/00224499.2016.1142495

Norrick, N. R. (2017). Language play in conversation. In N. Bell (Ed.), *Multiple perspectives on language play* (pp. 11–45). de Gruyter.

Ohlheiser, A. (2021, July 13). Welcome to TikTok's endless cycle of censorship and mistakes. *MIT Technology Review.* https://www.technologyreview.com/2021/07/13/1028401/tiktok-censorship-mistakes-glitches-apologies-endless-cycle

Pascual-Ferrá, P., Alperstein, N., Barnett, D. J., & Rimal, R. N. (2021). Toxicity and verbal aggression on social media: Polarized discourse on wearing face masks during the COVID-19 pandemic. *Big Data & Society*, January-June, 1–17. https://doi.org/10.1177/20539517211023533

Peterson-Salahuddin, C. (2022). "Pose": Examining moments of 'digital' dark sousveillance on TikTok. *New Media & Society*, 14614448221080480.

Rauchberg, J. S. (2022). #Shadowbanned: Queer, trans, and disabled creator responses to algorithmic oppression on TikTok. In P. Paromita (Ed.), *LGBTQ digital cultures: A global perspective* (pp. 196–209). Routledge.

Rettberg, J. (2017). Hand signs for lip-syncing: The emergence of a gestural language on Musical.ly as a video-based equivalent to emoji. *Social Media + Society, 3*(4), October-December 1-11 https://doi.org/10.1177/2056305117735751

Roberts, S. T. 2019. *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, USA, pp. 1668–1678. http://dx.doi.org/10.18653/v1/P19-1163

Schellewald, A. (2021). Communicative forms on TikTok: Perspectives from digital ethnography. *International Journal of Communication*, *15*, 1437–1457. https://ijoc.org/index.php/ijoc/article/view/16414

Sherblom-Woodard, B. (2002). *Hackers, gamers and lamers: The use of l33t in the computer sub-culture.* B.A. thesis, Swarthmore College. Institutional Scholarship. http://hdl.handle.net/10066/11233

Sherzer, J., & Webster, A. K. (2015). Speech play, verbal art, and linguistic anthropology. *Oxford Handbooks Online*. doi:10.1093/oxfordhb/9780199935345.013.33

Silverstein, M. (2003). Indexical order and the dialectics of sociolinguistic life. *Language and Communication*, *23*(3-4), 193-229. https://doi.org/10.1016/S0271-5309(03)00013-2

Sung, M. (2019, September 30). Trippy video breaks down how YouTube demonetizes LGBTQ creators for using words like "gay." *Mashable.* https://mashable.com/video/youtube-demonetize-lgbtq-words-nerdcity

TikTok. (2021, September 27). Thanks a billion! *TikTok Newsroom*. https://newsroom.tiktok.com/en-us/1-billion-people-on-tiktok

TikTok. (2022, February). *Community Guidelines*. https://www.tiktok.com/community-guidelines?lang=en#29

Valdman, A. (2000). The language of the inner and outer suburbs: From argot to popular French. *The French Review*, 73(6), 1179–1192.

Vásquez, C. (2019). *Language, creativity and humour online*. Routledge.

Wolfram, W., & Schilling, N. (2015). *American English: Dialects and variation*. Wiley.

Wright, K., Zimmer, B., Carson, C. E., Hughes, B., McLean, J. & Zhang, L. (eds.) (2023). "Among the New Words." *American Speech, 98*(3), 296-317. https://doi.org/10.1215/00031283-10887733

Zeng, J., & Valdovinos Kaye, D. B. (2022). From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet*, February, 1–17. https://onlinelibrary.wiley.com/doi/pdf/10.1002/poi3.287

Zimmer, B., Carson, C. E., Horn, L. R. (2011). Among the new words. *American Speech*, 86(3), 355–376. https://doi.org/10.1215/00031283-1503937

Zulli, D., & Zulli, D.J. (2020). Extending the internet meme: Conceptualizing technological mimesis and imitation publics on the TikTok platform. *New Media & Society*, 1–19. http://journals.sagepub.com/doi/10.1177/1461444820983603

Appendix

Categories for content moderation on TikTok as of March 2022

1. Minor safety
2. Dangerous acts and challenges
3. Suicide, self-harm, and disordered eating (formerly "Suicide, self-harm, and violent acts")
4. Adult nudity and sexual activities
5. Bullying and harassment
6. Hateful behavior (formerly "Hate speech and ideologies")
7. Violent extremism
8. Integrity and authenticity
9. Illegal activities and regulated goods
10. Violent and graphic content
11. Copyright and trademark infringement
12. Platform security
13. Ineligible for the For You Feed

Biographical Notes

Kendra Calhoun [kendrac@illinois.edu] is Assistant Professor of Linguistic Anthropology at the University of Illinois, Urbana-Champaign. Her research interests include social media discourse, Black digital communities, and the intersections of language, race, and power.

Alexia Fawcett [afawcett@ucsb.edu] is a doctoral candidate in Linguistics at the University of California, Santa Barbara. She researches the intersections of language and culture in discourse, morphosyntax, and gesture in online and offline contexts.