Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Dynamics of Language Change: A Mixed-Methods Analysis of Language in an Online Transgender Community

Cedar Brown, Lal Zimman, Simon Todd

Department of Lingusitics, University of California Santa Barbara
cedarbrown@ucsb.edu, zimman@ucsb.edu, sjtodd@ucsb.edu

Abstract

Trans-affirming language can be critical for trans people in transphobic sociopolitical contexts. But what kinds of trans activism – and whose – has determined what language is considered trans-affirming? Collective negotiations of trans language norms have long occurred across online platforms, providing rich data to explore what language change looks like in a community where language is highly politicized. In this paper, we use a mixed-methods approach to explore patterns of language change in one popular early online transmasculine community on LiveJournal. Using bigram snapshot language models, we establish canonical patterns of community-level change in aggregate language usage. We dig into how these patterns relate to population turnover by analyzing distinct groups of users, extending existing quantitative analysis to disentangle group-level variation from community-level change and combining it with qualitative analysis that scrutinizes the interactional contexts of change. We find influence of linguistic, interactional, and identity norms that are negotiated and enforced by highly active users, thus driving language change. Our analysis demonstrates the role of power in community-internal language negotiation, with implications for trans language change more broadly.

1 Introduction

For trans people, language can be a critical source of affirmation, with a growing body of research showing the negative impacts of misgendering on trans people's health and well-being (Ross, Kinitz, and Kia 2022).

As mainstream interest in trans-affirming language has grown, allies are often presented with a series of recommendations about how to respectfully refer to trans people. This picture, however, misses the far-ranging critiques trans people have put forth about language, including practices for referring to cis people and talking about bodies. It also omits how these recommendations arise from emerging changes negotiated within trans spaces. Given the importance of trans-affirming language, it is crucial to unpack how we arrive at such norms.

Since the beginning of public internet usage, trans people have engaged in collective negotiations of language norms. Such norms emerge from the sociopolitical contexts in which trans people exist, varied access to particular online platforms, and the dynamics of particular trans communities on those platforms (Dame-Griff 2019). One aspect of this is the socially and technologically specific ways that moderator[1] and community power are employed within such communities (Thach et al. 2024). To understand current norms, then, requires exploring earlier parts of trans language change and its context, including on platforms that are now largely defunct. While some early internet adopters used bulletin board systems or Usenet in the 1990s (Dame-Griff 2019), the early 2000s saw the popularization of blogging platforms with forums where users could interact. One such platform, LiveJournal, is described by Zimman and Hayworth (2020, p. 14) as "particularly popular among trans people in the 2000s" with "many of the discourses and terminological norms that have risen to some prominence... seen in early form here". This decade was also a time in which many trans communities shifted their attention toward promoting inclusion in mainstream gay and lesbian activism (for instance, in relation to attempts to pass the U.S. Employment Non-Discrimination Act) and pushing major NGOs like the Human Rights Campaign not to ignore trans issues (Valentine 2007). The popularity and relevance of LiveJournal makes it apt for studying the development of trans language norms online, both as a case-study to understand more about patterns of trans community language change generally and in itself as part of the discursive fabric that has contributed to current conceptions of trans language.

In this paper, we analyze language use in a popular transmasc LiveJournal community in the 2000s, as trans people began to connect over the internet in greater numbers. We take an exploratory mixed-methods approach to examine:

- **RQ1: How much is language use changing in an early online trans community?**
- **RQ2: What are the interactional dynamics surrounding early trans community language change?**

Our analysis comprises three parts. First, in Section 4, we

[1] When we discuss moderation, we refer to the moderation efforts our corpus makes visible: i.e., discussions of what language was and was not acceptable, rather than the deletion of posts or banning of users. Ethnographic insight suggests that the role of moderator and community member in this community was fluid, with frequent posters being recruited by the community as moderators.

analyze a specific lexical change as emblematic of changes occurring in the community. We find that the term *bio-* for non-trans individuals is gradually replaced by the term *cis-* – replicating the results of previous studies across a wider range of contexts. Second, in Section 5, we extend existing methodologies using bigram snapshot language models to assess how language norms in the community at large change over time. We find evidence for broader changes in language conventions on the same temporal trajectory as the lexical change, independent of variation due to changes in population size. Third, in Section 6, we zoom into three groups of users (frequent posters, new users, and leavers), pairing quantitative and qualitative methods to explore behavioral dynamics. Our analysis suggests a large role for metalinguistic debates and negotiation of norms, where community-internal language change is led by frequent posters, adopted by newcomers, and where being out of step with language change could factor into users leaving the community. This analysis highlights how certain individuals can have an outsized effect on language change in trans communities, denaturalizing current trans language norms in favor of seeing trans language as emerging in conversation with the way that gender societally unfolds. In the current highly politicized environment for trans people online (with, for example, social media giants Meta and X removing hate speech protections for trans people) trans people and trans-affirming language is often characterized by critics as monolythically imposing "out of touch" prescriptions (Conger 2025). This study instead presents a picture in which trans people have a wide diversity of opinions about language, with consensus emergent through years of metalinguistic discussion and debate. Within this, oppressive dynamics outside of trans communities are also seen to come into play within.

2 Past Approaches to Online Language Change

The study of language change in online communities has been approached in many ways. While each approach has its limitations, it also makes valuable contributions that we incorporate and build upon in the present work.

One approach examines the introduction and spread of innovative lexical items, identifying the characteristics of individuals that drive lexical change. For example, Del Tredici and Fernández (2018) showed that, across 20 different subreddits, the introduction and spread of new terms was driven by users who were most active and visible in the community or cliques within it. In our work, we incorporate this insight by asking whether users with different activity levels have different roles in language change. Distributional semantic approaches have yielding interesting insights into the way that terms shift in meaning over time (Schlechtweg et al. 2020; Soni, Klein, and Eisenstein 2021), an insight which we use to motivate our own incorporation of an embedding-based approach.

However, such approaches are limited in not considering changes beyond the lexical level. An alternative approach zooms out to examine community-wide changes in conventions of language use in aggregate, identifying how they relate to changes in individual users' language and in the composition of users in the community. For example, Danescu-Niculescu-Mizil et al. (2013) developed a framework that tracks changes in how statistically similar posts in a community are to each other, which incorporates influences of lexical and syntactic conventions as well as conventions in the kinds of topics that are discussed. They applied this framework to two online beer-rating communities and showed that, while a given user quickly adopted emerging conventions upon joining the community (thus advancing the spread of change), they did not continue following changes (such that community-level change was primarily driven by population turnover). This approach is taken up and modified by Törnberg and Törnberg (2024) in looking at how new users to a white power forum quickly converge to community language norms, with those who leave early being more divergent from these norms than those who remain. Our work further tests the generality of Danescu-Niculescu-Mizil et al. (2013)'s result by applying a similar framework to an online trans community – where conventions of language use have high personal and political stakes. In doing so, we extend the framework to incorporate the insight that separating newcomers and leavers can illuminate the effects of individual behavior and population turnover on community-level change in language conventions.

While quantitative approaches reveal aggregate patterns of language change, the have limited ability to uncover the way in which these patterns may depend upon social context within interactions. Qualitative approaches illuminate the interactional context of online language change (e.g., Thach et al. 2024), revealing reasons why a change occurs and the responses it engenders. Such studies explore how users construct community membership by employing similar linguistic strategies and community-specific terminology (Leuckert 2020; Rüdiger and Dayter 2022), as well as how posters and commenters take stances in relation to each other (Gordon and İkizoğlu 2017). Qualitative studies specifically looking at trans communities on Tumblr (Dame 2016; Jacobsen, Devor, and Hodge 2022) and YouTube (Miller 2019) have further explored the way that trans terminology is debated online. An example of such debate is seen in the trans LiveJournal community under study, Zimman (2014) details how the compound term *dic-clit* was met with objection due to its implicit assertion that trans men are not male-bodied, and how the shift to the un-compounded *dick* was linguistically empowering.In our work, we use qualitative analysis to understand the social context of the community, as reflected in the way that users behave, as well as the motivations and reactions to change, as revealed through metalinguistic discussions.

Purely qualitative approaches are difficult to apply at scale, and thus can miss aggregate community change that would be detected by quantitative approaches. This limitation is addressed by mixed-methods approaches, which combine quantitative and qualitative methods to give a comprehensive picture of language change. For example, using mixed-methods, Dame-Griff (2019) shows that the spread of the term *cisgender* in trans Usenet groups from 1992 was

driven by a few highly active users through the frequent posting of metalinguistic contestations that were extensively quoted in discussion and debate. This result shows that metalinguistic discussions not only provide a useful lens for the analysis of language change, but can also be heavily implicated in the unfolding of that change. In our work, we similarly complement the quantitative analysis of highly active users with qualitative consideration of the role that their commentaries on language and behavior played in the establishment and enforcement of community norms.

In this paper, we integrate insights and methods from a range of past approaches to study language change in an online transmasc community. This mixed-methods approach allows us to formulate a deep and contextualized understanding of the dynamics of language change in an online community of marginalized individuals, where language and identity are deeply intertwined.

3 Data

3.1 Data Source and Composition

The data for this paper come from the TransLiveCorpus (Zimman and Hayworth 2020), which contains message board interactions from four trans communities on LiveJournal.com that occurred between late 2000 and early 2018. Specifically, this paper focuses on the *ftm*[2] community, which was characterized by its users as the "largest and most visible" transmasc community on LiveJournal. This community was also selected as focus because of the community-insider knowledge of the research team, which includes two transmasc researchers, one of whom conducted participant-observation in the community during its most popular years. The data from the *ftm* community contains 219,711 entries[3] (19,631 posts and 200,080 comments) from 5,188 distinct users, comprising approximately 17 million words in total. These entries were posted throughout the span of the corpus, but only sporadically before 2002 and after 2012; therefore, we omit these periods from all analyses and figures.

As can be seen in Figure 1, the activity of the *ftm* community started rapidly increasing in 2002 and peaked in 2007 (see Appendix A for details). After this point, activity rapidly diminished, due to changes in both the number of active users and the frequency with which they posted. Most users were active and most entries were posted between 2003 and 2009.

3.2 Pre-processing

To prepare the data for analysis, we normalized and tokenized the text of each entry. To normalize, we used regular expression patterns to replace instances of smileys, e-mail addresses, phone numbers, URLs, and action strings (e.g., **falls off chair**) with generic category labels, so that they would not overinflate estimates of lexical variation. We also restored censored characters from common profanity terms (e.g., *sh!t*, *f*ck*), added start-of-sentence markers, replaced various end-of-sentence punctuation with a common marker, stripped word-external punctuation, and lowercased the text. To tokenize, we separated clitics (e.g., *'s*, *'d*) from their hosts and split the remaining text on whitespace.

To capture patterns of change over time, we binned the entries by the date on which they were posted. To balance representativeness (having enough entries in a bin to infer general rather than idiosyncratic properties through quantitative analysis) against temporal resolution (having enough bins to track the development of changes in a fine-grained manner), we chose to bin in 3-month periods (e.g., entries posted between January 1 and March 31 of a given year).[4]

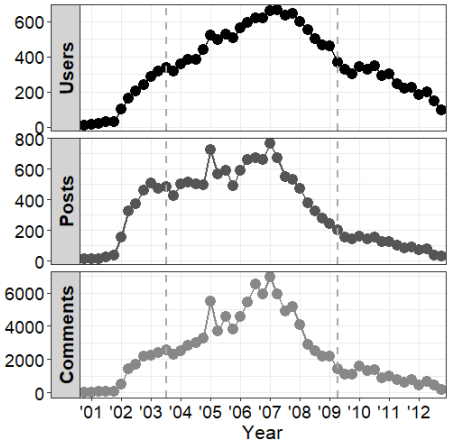

Figure 1: Number of active users (top), posts (middle), and comments (bottom) within each 3-month bin from 2001 through 2012, for the whole corpus. Activity rapidly increased over 2002, continued to gradually increase until a peak in 2007, and then rapidly decreased. Our core analyses focus on the period from mid-2003 through mid-2009, during which all measures maintained high values.

4 A Specific Lexical Change: *bio-* to *-cis*

The core of this paper (Sections 5 and 6) analyzes language at a large-scale quantitative level, investigating aggregate patterns of lexical and syntactic change. To contextualize and bolster understanding of such abstract changes, we begin by examining a single lexical change.

Over the lifetime of the *ftm* LiveJournal community, there were salient changes in identity terms (Zimman and Hayworth 2020). We examine a change in the words used to describe non-trans people, where the term *bio-* (as in *biomale*, short for 'biologically male') was replaced by *cis-* (as in *cismale*, with the Latin prefix *cis-*, the opposite of *trans-*, meaning 'on the same side'). The terms *bio-* and *cis-* show morphosyntactic flexibility: they can both be used as a prefix

[2] *ftm* stands for 'female-to-male', a term for people who don't identify with the female sex they were assigned at birth, but rather with maleness or masculinity. While it was once common, it is now disfavored due to its perceived binarity and bioessentialism; the current preferred term is *transmasc* (short for 'transmasculine').

[3] The term *entries* refers to both *posts* (standalone texts on the message board) and *comments* (threaded replied to the posts). Our analyses make no distinctions between posts and comments.

[4] We explored the suitability of bins of different sizes and found 3-months to be best; see Appendix B for discussion.

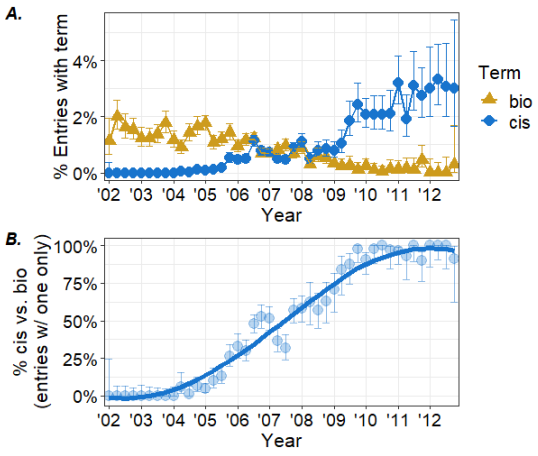

Figure 2: An example of a lexical change, from *bio-* to *cis-*. **A (top):** Entries in each 3-month bin containing *bio-* (yellow triangles) and/or *cis-* (blue circles), as a proportion of all entries in that bin; error bars show 95% binomial confidence intervals. **B (bottom):** Entries containing *cis-*, as a proportion of entries containing either *bio-* or *cis-* (but not both).

(*bioman*, *cis-men*) or an adjective (*cis guys*).

The change from *bio-* to *cis-* has previously been analyzed by Zimman and Hayworth (2020), who explored how frequently it modified a fixed set of person terms (e.g., *man*, *girl*, *people*). To examine this change in more detail, beyond this fixed set of terms, we extracted all words starting with *bio-* or *cis-* and all instances of standalone *bio* or *cis* followed by another word. For each case, we manually identified whether *bio-* or *cis-* was being used as either a prefix or adjective in a way that is relevant to trans identity.[5] We excluded non-modifying uses, where *bio-* or *cis-* was the syntactic subject or object, in order to limit the influence of metalinguistic discussions. In total, we identified 5,320 instances of *bio-*, of which 2,699 were relevant, and 2,175 instances of *cis-*, of which 2,058 were relevant. Relevant uses modified a wide range of lemmas, including many that weren't included in the previous analysis (see Appendix C).

We counted the number of entries in each 3-month bin containing *bio-* or *cis-*. Focusing on the number of entries allows us to see how the change is spreading through the community, separate from the question of how often individual users use either term in a given entry, and allows us to track the prevalence of these terms in general. To compare *bio-* and *cis-* as variants of a single lexical variable, we restricted focus to entries that contained only one of the terms; this restriction also further limited the influence of metalinguistic discussions, which often mentioned both terms.

As shown in Figure 2A, the use of *bio-* decreases and the use of *cis-* increases over time. When considering *bio-* and *cis-* as competing variants of a single lexical variable, the frequency with which one is used relative to the other traces out an S-shaped curve (Figure 2B), which is a canonical pattern of language change (Weinreich, Labov, and Herzog

[5]We did not treat the full form *biological* as relevant.

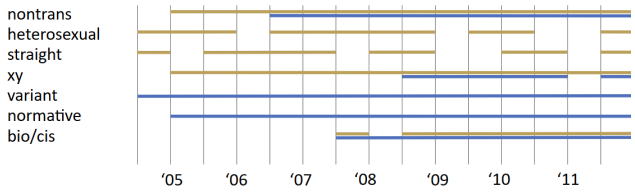

Figure 3: Semantic neighbors of *bio-* (yellow) and *cis-* (blue) over time, showing words that appeared in the top five nearest neighbors for either variant in ≥50% of time periods.

1968). Thus, there is a clear indication of language change in the *ftm* community over the period captured in our data, at least at the level of this single lexical variable. The period during which the change is most rapidly spreading through the community coincides with the period during which the community is most active, between 2006-2008. By around 2011, *cis-* is practically exclusively used in the community, and it appears to be used in a more widespread way (i.e., in a greater proportion of entries) than *bio-* ever was.[6] This result aligns with the ethnographic observation that *bio-* held a problematized status before *cis-* emerged as an alternative, evidenced through metalinguistic debate and moderation[1].

In the change from *bio-* to *cis-*, one term has not simply replaced another, because they do not modify the same lemmas in the same proportions (see Appendix C). Rather, the semantic domain in which the terms are used has also changed. To illustrate this, we trained embeddings[7] on all entries from each time bin and used them to identify and compare the top semantic neighbors of *bio-* of *cis-* over time. As can be seen in Figure 3, both terms relate to *nontrans*ness stably across time, but they do so in different ways. While *bio-* is highly similar to words about sexuality and chromosomes (*heterosexual*, *straight*, and *xy*), *cis-* is highly similar to words about social norms and conformity (*variant* and *normative*). Due to these differences, *bio-* and *cis-* are not highly similar to each other at first, and only become so as *cis-* comes to replace *bio-* as the general *nontrans* term. This shows how language change reflects change in community perpectives, which can be highly ideologically loaded (see Section 6.2 for discussion of associated tensions).

The change from *bio-* to *cis-* is just one example of language change within the *ftm* LiveJournal community, in a specific lexical variable. However, it provides expectations about what a typical trajectory of change might look like in the community, with temporal anchor points such as peak variation in 2007. We use these expectations to guide interpretation of more abstract indicators of language change. This serves as a source of insight on the dynamics of change

[6]The same patterns are seen when conducting the *bio-*/*cis-* analysis over users, based on which variant each user uses most in each time bin. This confirms that the overall change is observed throughout the community and not just in a few very active users.

[7]We used CBOW word2vec (Mikolov et al. 2013) to train embeddings, with window size 5 and embedding length 100. We used static embeddings (word2vec) rather than contextual ones (e.g., from BERT) because of data size constraints, and because they give us a single representation for each term (type) at each time bin.

in a community highly concerned with language.

5 Broader Changes in Language Use

Beyond the change from *bio-* to *cis-*, (and other similar changes in lexical variables such as the shift from *transgendered* to *transgender* explored by Zimman and Hayworth (2020)), there are likely broader changes in the conventions of language use in the *ftm* LiveJournal community. Such changes affect what entries typically 'look like', including conventions of which words are chosen for specific terms, how words are combined, and even which topics are open for discussion. To get a broader view of such changes, we turn to *snapshot language models* as a statistical representation of what language use in the community 'looks like', in aggregate, at different points of time (Danescu-Niculescu-Mizil et al. 2013).

Snapshot language models give us little ability to identify changes in specific lexical items, especially as compared to embedding-based alternatives (e.g., as reviewed by Schlechtweg et al. 2020). However, they are well suited to our goal of zooming out to get a holistic view of change patterns. Furthermore, because they focus at an aggregate level above individual lexical items, they are relatively robust to sampling noise related to the sparsity of data in each time bin.

5.1 Data: Filtering and Grouping

To facilitate the snapshot language model analysis, we conducted further filtering and grouping of our data.

We applied filters based on considerations of entries and users. For entries, we followed Danescu-Niculescu-Mizil et al. (2013) in filtering out entries containing fewer than 30 tokens. Such entries typically represent short responses to a preceding thread (e.g., "Awesome idea"), which are grounded in interactional contexts from other entries and are therefore not well-suited to language models that aim to capture the generation of streams of self-contained text. For users, we filtered out entries posted anonymously, as well as those from users who had fewer than 90 days (i.e., the span of a single time bin) between posting their first and last entry. This filtering ensured that each entry in our analysis could be linked to an identifiable user, where each user was active in the community for at least two (partial) time bins.

We then grouped users and entries in such a way as to enable consideration of the extent to which community-level changes reflect changes in the language of a given individual, versus changes in the composition of individuals within the community (Danescu-Niculescu-Mizil et al. 2013). These considerations are analogous to those made in sociolinguistic studies of language change in in-person communities, where groups are used to differentiate between changes at the individual and population levels and to examine effects of generational transfer (Weinreich, Labov, and Herzog 1968; Labov 2001; Tagliamonte and D'Arcy 2009).

We split our data into four groups, as summarized in Figure 4. Grounded in previous analysis of this community (Brown 2022; Zimman and Brown 2022), these groups were chosen to distinguish between users with different degrees

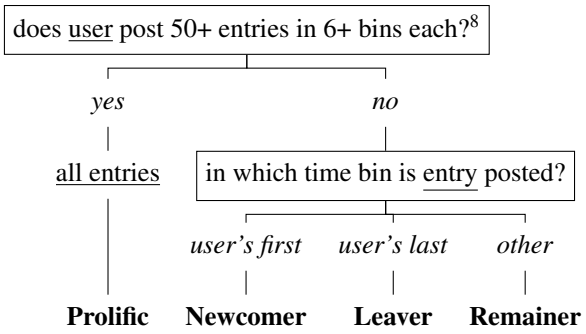

Figure 4: Decision tree for allocating users/entries to groups.

of engagement with the community and to facilitate consideration of population turnover. First, we separated the few most active (*prolific*) users from the 'everyday' regular users, to facilitate close-up investigation of how these most active users initiated or responded to changes.[8] Then, in each 3-month time bin, we split the entries of regular users between three groups depending on whether they had been posted during that user's first, last, or intervening 3-month time bin(s) in the community. The corresponding groups of *newcomer*, *leaver*, and *remainer* entries allow us to identify correlates of population turnover, following Danescu-Niculescu-Mizil et al. (2013), and to gauge the extent to which the linguistic changes we observe are specific to this LiveJournal community versus reflective of changes in the broader transmasc community.

We use remainers as our model of what language use in general 'looks like' in the community, and we use other groups as a way to investigate who may be driving, participating in, or responding to any community-level changes. Our analysis focuses on the period from mid-2003 through mid-2009, during which the community is most active in terms of both number of users posting and number of entries posted (see Appendix A for a breakdown of activity over time, and Appendix B for details guiding our selection of this time period). The total numbers of users, entries, and tokens for each group in our analysis are shown in Section 5.1.

5.2 Methods: Snapshot Language Models

To characterize what general language use in the community 'looks like' at various points of time, we trained *snapshot language models* on entries posted by remainers in each 3-month bin. Our models were trained using the

[8]The prolifics group comprises 13 users who posted a substantive entry (consisting of at least 30 tokens) at least every other day, on average, for over 1.5 years. This threshold was determined through inspection of the distribution of entries per time bin across users (see Appendix A). In order to develop an in-depth profile of prolific users, we treated 'prolificness' as a static attribute: all entries from a prolific user were included in the prolifics group, even if they occurred before the user first met the threshold or after they stopped meeting it. This treatment also ensures that commonalities between the prolifics group and other groups at different points of time in the quantitative analysis cannot be due to the idiosyncracies of a user that appears in both groups at some point.

Group	Users	Entries	Tokens
Prolifics	13	11616	1390460
Newcomers	1737	12591	1505241
Leavers	1384	5112	590830
Remainers	1803	77549	8749379

Table 1: Total users, entries, and tokens included in our analysis for each group between mid-2003 and mid-2009.

SRILM toolkit (Stolcke 2002), to the same specifications as those used by Danescu-Niculescu-Mizil et al. (2013): they were bigram language models with Katz backoff smoothing, where the unigram back-off distribution was smoothed with Laplace (additive) smoothing with a smoothing parameter of 0.2.[9] A snapshot language model from a given time bin represents an estimated statistical summary of language use by 'everyday' users (remainers) during that bin, based on consideration of which words were used, how often they were used, and how they were combined with each other.

Following Danescu-Niculescu-Mizil et al. (2013), we measured how adequately each snapshot language model from a given time bin statistically predicts the language used by remainers in test samples of contemporaneous entries that it was not trained on, and we tracked changes in this metric over time as an indictator of language change. Danescu-Niculescu-Mizil et al. (2013) used cross-entropy for this purpose; we extended this approach by decomposing cross-entropy into entropy and Kullback-Leibler divergence. This allows us to disentangle whether difficulty in statistically predicting the language used in a test sample is because it has high inherent variability (high entropy) or because it is inconsistent with the patterns of language use captured by the snapshot language model (high divergence). In this way, we can interpret changes in idiosyncratic variation associated with population dynamics separately from changes in community-wide conventions. For the technical details of calculating entropy and divergence, see Appendix D.

We took several steps to ensure consistency in the representativeness of analyses across time bins. First, we used the same amount of training and test data for each model: 30 tokens per entry, with 500 entries for training and 100 for test.[10] This ensured that all analyses had the same level of granularity, regardless of how much activity there was in the corresponding bin. Second, we used a fixed quota for the number of users and entries per user in the training data.[11] This ensured that no user was over-represented in the training data, and that the extent of representation of different users was held constant for all models. Third, we took the 30 tokens per entry evenly from throughout the entry: we took the first, middle, and last 10-token snippets.[12] This provided even coverage of parts of the entry that may be fulfilling distinct functions, such as responding to earlier comments in a thread, making self-contained statements, and summarizing / signing off. Finally, we trained 100 snapshot language models for each time bin, each based on a different random sample of entries from remainers in that bin (subject to the above constraints), and tested each one on 100 different random test samples. By aggregating our analysis over many models and test samples for each bin, we ensure that the results are not driven by the specific composition of the training or test data, but are rather representative of how all of the 'everyday' users used language in that bin.

Both entropy and divergence measure linguistic variation. Entropy measures variation in the test set; high entropy indicates that the entries in the test set use a wide variety of linguistic units, with limited re-use across entries, such as may be expected when the test set is composed of entries from many users posting on diverse topics. Divergence measures variation between the training and test sets; high divergence indicates that the entries in the test set differ in their use of linguistic units compared to expectations formed from the training set, such as many be expected when there are distinct conventions of language use. For present purposes, we are most interested in changes in divergence over time, as a measure of changes in self-similarity that may reflect the establishment or abandonment of conventions of language use (cf. Danescu-Niculescu-Mizil et al. 2013), separate from changes in entropy that may reflect population dynamics.

5.3 Results

We measure language change by tracking entropy and divergence across time bins. In a situation where conventions of language use follow an S-shaped change, we expect to see divergence rise and then fall, yielding an inverted U-shape over time. If the specific instance of S-shaped change that we observed in Figure 2 is part of a broader change in language use in the *ftm* LiveJournal community, we may expect a similar temporal trajectory in overall divergence, with a peak around 2007.[13] To the extent that such a change is not merely a reflection of population growth and decline, we expect entropy to show a different trajectory.

The results are shown in Figure 5. We find an inverted U-shaped trajectory for divergence with the same tempo-

[9] We chose classical bigram language models over modern NN-based alternatives because their high bias is beneficial in limiting overfitting to our small training sets (15,000 tokens/model), because they can be trained entirely on our data without requiring pretraining on other data that may contaminate the analysis, and because they facilitate comparison with previous work.

[10] Our choices follow Danescu-Niculescu-Mizil et al. (2013), who used 30 tokens per entry and trained models on 1000 entries. Given limitations of dataset size, restricting training to 500 entries per time bin allows us to maximize the time span we can analyze.

[11] These choices are inspired by Danescu-Niculescu-Mizil et al. (2013), who used 2 entries from each of 500 users. Due to the distribution of entries over users across time bins (see Appendix B), the 500 training entries in our analysis were composed of 3 entries from each of 113 users, 2 entries from each of 53 different users, and 1 entry from each of 55 yet different users.

[12] For language modeling, we represented the first token in the first snippet as a bigram with a start-of-sentence context, and the first token in the middle and last snippets as a unigram.

[13] Since our language model analysis ends in 2009 due to data size limitations, we do not expect the inverted U-shape to be symmetrical; while we expect to see a decrease after 2007, we do not expect divergence in 2009 to be as low as it was in early time bins.

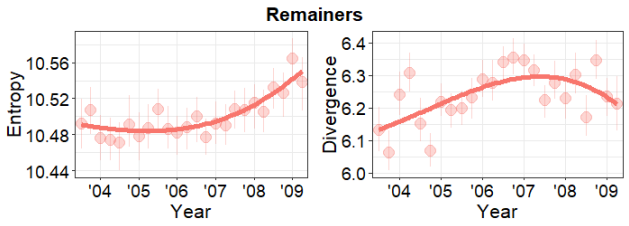

Figure 5: Entropy and divergence for samples of 100 entries from remainers, when predicted by a bigram language model trained on other remainer entries in that time bin. Error bars show 95% percentile confidence intervals of the sample mean and cubic smoother shows the general trend.

ral dynamics as the lexical change from *bio-* to *cis-*, supporting the idea that the lexical change reflects a broader change in language use in the community. The trajectory for entropy is distinct, suggesting that this broader change involves turnover between conventions of language use, rather than just increases and decreases in linguistic variation as the population of everyday users grew and declined.

6 Understanding Change: User Groups

To better understand the apparent broad pattern of language change discussed in Section 5, including who may be driving or responding to it, and how it may be interacting with the composition of the community, we turn to consideration of different groups of users. For each non-remainer group described in Section 5.1 – prolifics, leavers, and newcomers – we use snapshot language models to investigate how their language usage compares to that of the remainers in which the change pattern was observed. We support the quantitative analysis of each group with insights from qualitative analysis that contextualize the linguistic and interpersonal behavior of that group over time, revealing motivations and mechanisms of the changes that we interpret.

6.1 Methods: User Group Analysis

Quantitative Methods Our analysis examines the extent to which the snapshot language models described in Section 5.2 statistically predict the entries posted by prolifics, leavers, and newcomers in each time bin. This gives us a measurement of how the similarity of language use between each of these groups and the remainers changes over time.

In each time bin, for each group, we formed 100 test samples of 100 entries each (where each entry was again represented by its first, middle, and last 10-unit snippets). We calculated entropy and divergence for each test sample, using each of the 100 snapshot language models trained on remainer entries. By averaging across test samples and language models within each time bin, we obtained estimates of entropy and divergence that control for sampling variation.

For each group, we track entropy and divergence from the remainers across time bins. Increasing entropy over time indicates increasing linguistic variation in that group, while increasing divergence indicates a greater separation in the linguistic conventions used by that group and the remainers. Again, our primary interest is in divergence: if a group is participating in the changes in conventions of language use previously evidenced among remainers, we expect their divergence to be approximately constant over time.

Qualitative Methods To contextualize and better understand the quantitative change patterns, we qualitatively analyzed entries from each non-remainer group based on our previous in-depth analysis of similar user groups (Brown 2022; Zimman and Brown 2022).

For each group, we performed targeted analysis of entries containing key terms that were relevant to our research goals, following the methodology of Zimman (2014). For example, we focused on metalinguistic and community-reflexive commentary by extracting entries that include the terms identified by Zimman and Hayworth (2020) as participating in change (such as *bio-* and *cis-*), as well as entries that contain terms related to community language norms such as *norm*, *language*, and *community*. We additionally analyzed entries for each group based on criteria specific to the group; for example, for prolifics, we considered all entries posted by each individual to track their behavior over time, and for newcomers, we analyzed posts tagged as 'introductions' to assess initial language use when first joining the community. Two co-authors with insider perspectives analyzed a random sample of such entries, with all authors conferring on central and problematic cases. We used thematic coding to identify broader themes, as well as discourse analysis to examine the way users utilized language to take stances in comment threads (Du Bois 2007). Such qualitative analysis allows exploration of how meaning is negotiated through interaction, attending to the (inter)personal stakes of language use. It thus allows us to assess various interpretations of the quantitative patterning of different groups. Further, targeted discourse analysis allows attention to language used by and about marginalized subgroups that may be invisibilized by a more aggregate analysis.

6.2 Results: Prolifics

The results for prolifics are shown in Figure 6. Prior to mid-2005, we observe an increase in entropy accompanied by a decrease in divergence, indicating an increase in linguistic variation that makes the prolifics more similar to remainers. This aligns with changes in the composition of the group (see Appendix A). Early on, the prolific group was made up of a few users who were instrumental in establishing the community. They acted as both ambassadors and moderators[1]: they posted entries that welcomed newcomers and offered advice, as well as entries that attempted (successfully or unsuccessfully) to shape and enforce community norms. Later on, the prolific group expanded to a wider variety of members who had joined and become heavily involved in the community after it started growing, some of whom took on ambassadorial and moderator roles less often and instead participated more in 'everyday' discussions. The initial period therefore reflects a dilution of roles and ideas about the community as a wider variety of users became represented among prolifics.

After mid-2005, once the composition of the prolifics sta-

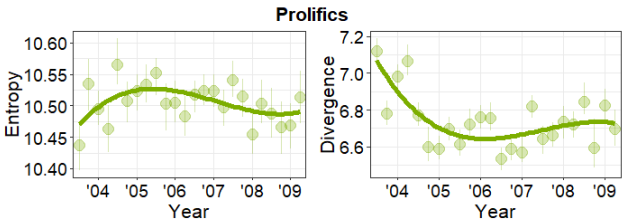

Figure 6: Entropy and divergence for samples of 100 entries from prolifics, when predicted by a bigram language model trained on entries from the remainers in each time bin. Error bars show 95% percentile confidence intervals of the sample mean and cubic smoother shows the general trend.

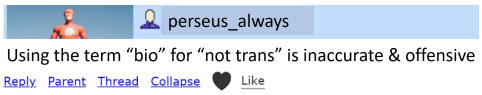

Figure 7: An example of metalinguistic correction by prolific user perseus_always (a pseudo-username)

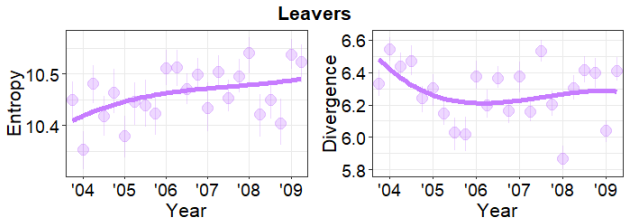

Figure 8: Entropy and divergence for samples of 100 entries from leavers, when predicted by bigram language models trained on entries from remainers.

bilized, we observe a slow but steady rise in divergence, indicating that the language of prolifics is becoming less similar to that of remainers over time. One possible explanation for this is that prolifics fall behind in the change exhibited by remainers, which would be consistent with demonstrations in other online communities that an individual's linguistic conventions may not change much once they are established (Danescu-Niculescu-Mizil et al. 2013). However, this is unlikely, as the concurrent decrease in entropy indicates that prolifics show less linguistic variation over time, implying movement toward a specific set of linguistic conventions.

To understand the post-2005 patterns, it is important to recognize that linguistic and interactional norms were constantly (re)negotiated as the ftm community expanded. As the most visible users, prolifics provided a model of linguistic and interactional behavior that others are likely to have implicitly adopted. In addition, those prolifics that took on community- and self-appointed moderator roles (formally and informally) were heavily involved in contentious discussions ("arguing back-and-forth", in the words of users) that aimed to explicitly identify appropriate norms (or "community rules"). Through their strong participation in these discussions, and subsequent enforcement of the norms they yielded, prolifics exerted substantial influence on the way that community members could use language.

For example, prolifics were instrumental in the change from *bio-* to *cis-*. Throughout the corpus, prolifics can be seen repeatedly objecting to the way that *bio-* naturalizes non-trans identities and implies that the maleness of trans men is somehow not biological, overriding counter-objections from other users that *cis-* is non-transparent and indexes academic class elitism. Through long explanatory posts and short forceful corrections such as the one shown in Figure 7, prolifics appear to have persuaded other users to use the term they deemed appropriate. They also intervened in discussions on uses of other slurs such as the *n-word* and the ableist *r-slur* where members lacked consensus on whether these words were acceptable. Additionally, prolifics were instrumental in establishing and enforcing interactional norms. For example, one prolific in mid-2003 asserted their right to post on any topic, but changed their position within a year when they replied to another user's post by stating "we *do* encourage people to keep posts specifically about ftm-related issues . . . you'll probably get better responses from more people if you stay on topic." This prolific not only rapidly adapted to a change in interactional norms, but also accelerated the change in others by enforcing the new norm. Because constraints on topic cause certain kinds of words and sentences to be avoided, this also translates into the prolific influencing community language use.

Based on the totality of this evidence, we interpret the increase in divergence and decrease in entropy after 2005 as indicating that prolifics drove the community-level change in which remainers participate; i.e., that they underwent the same changes as remainers, on a faster timescale. A logistic regression analysis further supports this interpretation, showing that prolifics are significantly ahead of remainers in the change from *bio-* to *cis-* explored in Section 4 ($\beta = 0.94$, 95% CI: $[0.61, 1.28]$).[14]

6.3 Results: Leavers

The results for leavers are shown in Figure 8. Entropy increases at first, indicating that users who leave during the first few time bins exhibit less linguistic variation than users who leave later. Similarly, divergence decreases at first, indicating that these initial leavers are also more linguistically distinct from the remainers. Given preceding evidence that these initial periods of analysis reflect the onset of language change in the community, we interpret these patterns to suggest that leavers are linguistically conservative; i.e., that they are those who are least inclined to participate in the incipient change. As language change progresses, divergence starts to increase, which suggests that users who begin to follow the change but don't complete it are likely to leave when their language, too, becomes outdated.

The relationship between using outdated language and leaving emerges throughout the time span in our qualitative

[14] In real terms, prolifics have a lead of approximately 1 year (12.6 months; 95% CI: 8.1–17.1 months) in the change from *bio-* to *cis-*, since the coefficient for a single 3-month time bin is $\beta = 0.22$.

analysis. Enforcement of norms (as in Figure 7 and Section 6.2) could contribute to participants leaving, as assertions of linguistic norms were often contentious. Looking at their final period in the community, we see leavers taking critical stances towards the calling out of certain language as "offensive" (with, for example, one 2007 leaver arguing that being sexist, racist, and offensive, shouldn't mean that someone should be "excluded from...getting rare FTM conversation and answers") and discussions of how identity terms such as *trans* and *ftm* should be defined. One particularly contentious linguistic debate that arose repeatedly was whether it is appropriate for trans men to use the word *bitch*, which some vocal members saw as unacceptably misogynistic.

Many of these leavers discussed the role of community moderation in creating and enforcing norms. For example, one 2006 post claimed moderators "have zero problem obliterating non-offensive posts that are obviously relevant to any normal human". Others discussed a more general culture of critique, with one 2007 leaver's post stating that "a lot of the people here... are on the edge of their seats waiting to fight over the wording of one sentence in a 700-word post." In these statements, leavers negatively evaluate moderators and critical users as ab-"normal" and desperate to fight. Some leavers discussed how a focus on language could exclude those outside of academic circles, while others discussed how posts perceived as "off-topic" could incur "insults." In this context, some leavers described a negative affective response to the very idea of posting, with a user in 2006 saying: "people are afraid to post. they reword entries zillions of times for fear of angering the entire community." By using the collective noun "people", this user legitimizes their affective stance, reducing its subjectivity by positioning it as shared by many. Because such commentaries are posted shortly before these users leave, they can be interpreted as outlining a motivation for leaving: leavers feel unwelcome and poorly aligned with the community, especially due to being moderated for using language that is at odds with the group's established norms. In fact, an offshoot community was created in 2007 called *free_speech_ftm*, which was described by its creator in a post to *ftm* as a place where there would be "no censorship of language, ever." In this, leavers' criticism of characterizing language as "offensive" locates the problem in an individual taking "offense" to language deemed racist, sexist, ableist, and fatphobic (among other things), rather than in the language itself and the underlying normative assumptions it conveys.

6.4 Results: Newcomers

The results for newcomers are shown in Figure 9. We observe an increase in entropy over time, indicating increases in linguistic variation and implying that users with a wider range of language practices may be joining the community during times of peak activity and turnover in linguistic conventions. We also observe near-constant but slightly-increasing divergence ($\beta = 3.0 \times 10^{-3}$, 95% CI: $[0.4, 5.9] \times 10^{-3}$ by linear regression), suggesting that newcomers are participating in linguistic changes in almost the same way as existing community members, but may lag behind a little.

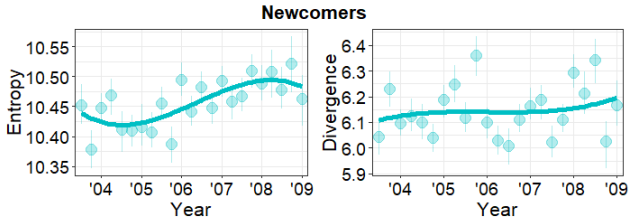

Figure 9: Entropy and divergence for samples of 100 entries from newcomers, when predicted by bigram language models trained on entries from remainers.

One possible interpretation of this pattern is that the language changes evidenced in the *ftm* LiveJournal community are not specific to this community, but are rather slightly-accelerated versions of broader changes in online language usage among trans individuals. An alternative interpretation is that newcomers are highly receptive, such that they observe the range of variation in linguistic conventions of the community prior to making their first posts, and rapidly situate their own language use within that range. Qualitative examination of the ways newcomers position themselves with respect to the community in their initial posts supports the latter interpretation. It was not uncommon for users to admit "lurking for some time" (2003) or "reading the posts... for quite a while" (2004) prior to posting or even having a LiveJournal account, suggesting that newcomers were adapting to changes within the community prior to posting. Further, while some newcomers remarked that they had been present in other trans communities online (e.g., LiveJournal's *genderqueer* forum) or in-person (e.g., in Buffalo), others indicated that they had little experience with trans communities, meaning they could not have picked up trans-specific terminology elsewhere. Such users had strong motivations to fit in with the first group of trans people with whom they had contact, which they may have accomplished (in part) by adopting community linguistic conventions.

7 Discussion

7.1 Trans Language Change

Our first research question asked how much language use was changing in an early online trans community. The quantitative results we observe are notably different to those observed in the context of the online beer-rating communities studied by Danescu-Niculescu-Mizil et al. (2013). These differences could reflect a number of key distinctions, such as differences in population dynamics, platforms, and the personal and political stakes of language in the community. Informed by qualitative analysis, our interpretation focuses on the last of these distinctions.

Language has very high stakes in the *ftm* community (and the trans community broadly) due to its potential to publicly represent a core identity aspect of marginalized individuals on their own terms. Consequently, language may be subject to more metalinguistic scrutiny in this community than in others (Zimman 2014; Dame-Griff 2019), especially in response to evolving societal norms and institutional poli-

cies. The sociopolitical precarity of many trans people could drive efforts to enact large-scale change in language conventions so as to better legitimate trans identities or more fully characterize users' lived realities, with strong pressure for community members to collectively define, use, and enforce new conventions. Indeed, we see whole threads in the *ftm* community debating specific lexical changes in trans terminology, with many users explicitly invested in a project of community consensus around transmasculine language use. Such metalinguistic commentary and correction, as with the illustrated change from *bio-* to *cis-*, may cause those using other terminology to change their language or leave the community (for example, to the alternative *free_speech_ftm* community). These dynamics illustrate the complex politics of enforcing language norms, wherein certain community members become marginalized in an effort by some to gain traction and legitimacy in mainstream conversations about rights, such as employment non-discrimination.

By modelling bigrams, we capture both the terminology itself and the way that people are putting language together, allowing us to zoom out to broader changes in the combination of word use and interactional patterns. Our qualitative analysis shows that interactional norms were also a source of metalinguistic commentary and contention. With the increasing connectivity of many geographically dispersed trans people throughout the 2000s, the way that people engaged with the community shifted. We provide examples of such shifts in topics of conversation (e.g., a prolific user shifting from discussing a range of topics, to encouraging others to keep topics specifically trans-related) and in modes of interaction (suggested by metalinguistic commentary describing an increasingly antagonistic interactional environment). While we are not able to tease apart these factors on the basis of our current analysis, our quantitative and qualitative findings suggest a broad shift in lexical and interactional norms in tandem with broader societal changes. Significant changes were seen even in a short span of time, suggesting that language change in a digital space is highly responsive to sociopolitical and technical factors.

7.2 Power, Contestation, and Moderation

Our second research question asks what the interactional dynamics surrounding early trans community language change are. Motivated by past research, we facilitated deeper understanding of language change in the *ftm* community by zooming in on three groups of users: the 13 *prolifics* who were most active across time, as well as the *leavers* of and *newcomers* to the community in each time bin. Our quantitative and qualitative analyses suggest that prolifics can drive change, in part through taking on moderator roles where they determine and enforce norms of language use, topic, and interaction. While the importance of interrogating formal moderator power has been shown in previous studies (Thach et al. 2024), in this community, we observed that the line between formal and informal moderation was blurry, with many without formal moderator roles explicitly commenting on appropriate language use, and the exacting of more severe kinds of moderator power (such as banning users) rare. These terminological enforcements can cause friction with individuals who oppose or lag behind community-level changes, ultimately driving them to leave the community. We saw this with the creation of *free_speech_ftm*, which depicted not moderators but the entire community as engaging in "censoring" people who used misogynistic, ableist, or racist language. These dynamics make explicit the interactional and linguistic conventions that newcomers can use to rapidly adapt when they first join the community.

While interactional dynamics may affect users participating in the community, these users also guard the possibilities of trans community – and even the boundaries of transness – for those who may "lurk" in the community without engaging (see Section 6.4). For such lurkers, the metalinguistic contention reveals gaps between who is *ideally* welcome in the community and who is *actually* welcome. For example, efforts to create a community in which everyone feels welcome by discouraging terms deemed racist, ableist, and/or sexist (such as the *n-word*, *r-slur*, and *bitch*) were met with opposition. This suggests that discrimination reflected by the use of such terms was normalized in the community (mirroring criticism of generic LGBTQ+ spaces as being overwhelmingly white and ableist; see e.g. Fox and Ore 2010), an idea which we have explored in-depth in previous work (Brown 2022; Zimman and Brown 2022). This may limit access to such a community, despite its crucial role in sharing peer information about medical and logistical aspects of transition and fulfilling key mental health needs. Further exploration of such dynamics and their implications for access are important topics for future work.

Our analysis of leavers and newcomers is in line with Danescu-Niculescu-Mizil et al. (2013)'s results, where users initially participate in ongoing language change and leave when their language is most conservative. By including the prolific group, we show the large role played by those with community presence and power in language change. As such, our analysis expands upon the connection between membership turnover and language change by demonstrating how users at different stages of community participation respond to changes encouraged by moderators and other critical users. Methodologically, it underscores the importance of interrogating the interactional way that change unfolds in a specific community, to capture the influences of local and contextual dynamics of power that can easily be missed when aggregating data across multiple communities.

7.3 The Platform

The community's spatiotemporal positioning on LiveJournal in the 2000s impacts the changes we see and their broader significance in two key ways.

First, the accessibility of LiveJournal increased the scale at which conversations about trans language were happening. While, for example, discussions of *bio-* and *cis-* were happening earlier (e.g., Dame-Griff 2019), these discussions were the purview of a few digitally-savvy people. LiveJournal saw the confluence of thousands of trans people, providing a fertile space for language changes to occur, and the *ftm* community was by far the most popular trans group on the platform (comprising 72% of a corpus of 4 communities

considered by Zimman and Hayworth 2020).

Second, the *ftm* community set precedents for continued consideration of linguistic standards in later online trans communities. Part of this is due to personal influence: some members of *ftm* likely shifted to emerging platforms like tumblr and Facebook as LiveJournal activity declined, spreading established linguistic norms to new audiences. But, a large part of the precedent is due to the accessibility of LiveJournal and *ftm*'s status as a public forum, which means that the community-internal negotiation and ratification of norms survives through an enduring written record that remains available to individuals searching for information about trans language. Due to the precedent, the leaders of linguistic change within *ftm* may have had have an outsized effect not only contemporaneously on the *ftm* community, but also consequently on the broader trans community. An important aspect of this is that, even though the LiveJournal *ftm* community was an open community without a geographically-defined name, it was largely US-centric in its discourse and membership. The US-centrism of generic online trans groups does symbolic work in contributing to a perception of transness or queerness as a US, English-centered phenomenon. Such discourses are able to be co-opted by conservatives in other countries to invalidate local understandings of gender diversity and efforts to address queer oppression (e.g., Borba 2019; Tudisco 2021). In this, we can see the way that online trans spaces are not only influenced by the wider sociopolitical climate, but also shape how social categories are understood.

8 Conclusion

Our mixed-method analysis demonstrates the impact of moderation and the enforcement of norms on language change in an online trans community. We first established the existence and timescale of change through exploring an example of lexical change (the shift from *bio-* to *cis-*) which exhibited an S-shaped curve typical of language change. Through an embeddings-based approach, we saw that this change was not only a replacement of terminology, but a shift in domains of discussion. We then zoomed out to consider aggregate language usage, based on snapshot language models over posts and comments of 'everyday' users, which showed an inverted U-shape in divergence with the same temporal trajectory as the lexical change, suggesting broader changes in language conventions. Finally, we explored the role that different kinds of users had in language change, by pairing quantitative analysis of group-wise entropy and divergence with qualitative analysis of metalinguistic commentary. Together, these analyses suggest that language change was driven by users who post frequently, followed by newcomers to the community, and contributed to some users ultimately leaving the community, highlighting the role of moderating metalinguistic commentary in community change.

Our methodology expands previous work in two key ways: (1) we extended the framework presented by Danescu-Niculescu-Mizil et al. (2013) by breaking down cross-entropy into entropy and divergence, allowing us to disentangle sources of linguistic variation when comparing specific user groups to the community at large; (2) we separated out prolific users and paired quantitative with qualitative analysis to more clearly understand the interactional dynamics underpinning aggregate-level changes.

Our analysis encourages attention to the historical contingency of current trans language norms and the sociotechnical environments that have shaped them. Given the crucial role trans-affirming language can play in trans people's health and well-being, it is important to attend to how online community-internal power-dynamics can contribute to broader conceptualizations of trans language. By providing a more nuanced picture of the development of norms, we open up space for the language practices of more marginalized trans subjects, as well as demystifying a complex process of linguistic convergence that might be resisted outside of trans communities because of its opacity.

9 Limitations

This paper has limitations of data, analysis, and scope.

Our data presented limitations in terms of messiness and sparsity. First, the original corpus stripped newline characters, which limited our ability to robustly identify some sentence boundaries, and contained typos and spelling variations that introduced noise. Second, we had to bin data in 3-month time bins in order to have enough entries to train each snapshot language model, but still did not have a large number of entries to model for early and late bins. This limited our ability to see fine-grained temporal dynamics, to generalize beyond bigrams, and to quantitatively examine language change beyond the 2003-2009 window. Third, our quantitative conclusions are based on a subset of the language in the community as, in order to control for length effects, we excluded shorter posts and portions of longer posts.

Our method of analysis presented limitations in interpretability. First, snapshot language models provide a bird's-eye view of consistency in language use across entries and user groups, which drastically simplifies the rich complexity of language variation and change – not permitting close quantitative examination of precisely *what* is changing, nor of *how* change spreads. Second, our quantitative observations are correlational, not causal. By piecing together various quantitative results, we have suggested interpretations about causes and responses to language change, which are supported by our qualitative analysis; however, alternative interpretations remain possible, as suggested in Sections 6.2 to 6.4. Third, our prolifics analysis is based on just 13 users, whose idiosyncracies may have affected our interpretation. Future work could use alternative methods to zoom in on specific changes and/or social networks.

Finally, as this is just one community on one platform during one period of time, we are limited in our ability to generalize. As the forum is in English and largely US-centric, our conclusions may not extend to other languages and cultures. Nevertheless, the fact that most forums of this nature are in English raises interesting points about the construction of trans community, which we discuss in Section 7.3. Future work is needed to explore whether the patterns we observed could be seen in other communities, platforms, and times.

Acknowledgements

Thank you to the work of Will Hayworth and Lal Zimman in creating the corpus this paper used. Thank you also to the CPLS lab, CEiLing research group, and Trill lab at UCSB for their feedback and encouragement. Finally, thank you to Mattia Samory and the ICWSM reviewers for their helpful comments.

References

Borba, R. 2019. Gendered Politics of Enmity: Language Ideologies and Social Polarisation in Brazil. *Gender and Language*, 13(4).

Brown, C. 2022. The Politics of Community Language Change: A Computational Analysis of Language Norms in an Online Trans Community.

Conger, K. 2025. Meta Drops Rules Protecting L.G.B.T.Q. Community as Part of Content Moderation Overhaul. *New York Times*.

Dame, A. 2016. Making a Name for Yourself: Tagging as Transgender Ontological Practice on Tumblr. *Critical Studies in Media Communication*, 33(1): 23–37.

Dame-Griff, A. 2019. Herding the 'Performing Elephants': Using Computational Methods to Study Usenet. *Internet Histories*, 3(3): 223–244.

Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities. In *Proceedings of the 22nd International Conference on World Wide Web*, 307–318.

Del Tredici, M.; and Fernández, R. 2018. The Road to Success: Assessing the Fate of Linguistic Innovations in Online Communities. In *Proceedings of the 27th International Conference on Computational Linguistics*, 1591–1603.

Du Bois, J. W. 2007. *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*. Amsterdam: John Benjamins.

Fox, C. O.; and Ore, E., Tracy. 2010. (Un) Covering Normalized Gender and Race Subjectivities in LGBT "Safe Spaces". *Feminist Studies*, 36(3): 629–649.

Gordon, C.; and İkizoğlu, D. 2017. 'Asking for Another' Online: Membership Categorization and Identity Construction on a Food and Nutrition Discussion Board. *Discourse Studies*, 19(3): 253–271.

Jacobsen, K.; Devor, A.; and Hodge, E. 2022. Who Counts as Trans? A Critical Discourse Analysis of Trans Tumblr Posts. *Journal of Communication Inquiry*, 46(1): 60–81.

Labov, W. 2001. *Principles of Linguistic Change, Vol. 2: Social Factors*. Oxford: Blackwell.

Leuckert, S. 2020. Rethinking Community in Linguistics: Language and Community in the Digital Age. In Jansen, B., ed., *Rethinking Community through Transdisciplinary Research*, 111–125. Palgrave Macmillan.

Mikolov, T.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the 1st International Conference on Learning Representations*.

Miller, J. F. 2019. YouTube as a Site of Counternarratives to Transnormativity. *Journal of Homosexuality*, 66(6): 815–837.

Ross, L. E.; Kinitz, D. J.; and Kia, H. 2022. Pronouns are a Public Health Issue. *American Journal of Public Health*, 112(3): 360–362.

Rüdiger, D.; and Dayter, S. 2022. *The Language of Pick-Up Artists: Online Discourses of the Seduction Industry*. New York, NY: Routledge.

Schlechtweg, D.; McGillivray, B.; Hengchen, S.; Dubossarsky, H.; and Tahmasebi, N. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 1–23.

Soni, S.; Klein, L. F.; and Eisenstein, J. 2021. Abolitionist Networks: Modeling Language Change in Nineteenth-Century Activist Newspapers. *Journal of Cultural Analytics*, 6(1): 1–43.

Stolcke, A. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proceedings of the Seventh International Conference on Spoken Language Processing*, 901–904.

Tagliamonte, S. A.; and D'Arcy, A. 2009. Peaks Beyond Phonology: Adolescence, Incrementation, and Language Change. *Language*, 58–108.

Thach, H.; Mayworm, S.; Delmonaco, D.; and Haimson, O. 2024. (In)visible Moderation: A Digital Ethnography of Marginalized Users and Content Moderation on Twitch and Reddit. *New Media & Society*, 26(7): 4034–4055.

Törnberg, P.; and Törnberg, A. 2024. Inside a White Power Echo Chamber: Why Fringe Digital Spaces are Polarizing Politics. *New Media & Society*, 26(8): 4511–4533.

Tudisco, J. 2021. Queering the French Académie: Reclaiming Linguistic Authority for Trans and Non-Binary People. *Toronto Working Papers in Linguistics*, 43(1).

Valentine, D. 2007. *Imagining Transgender: An Ethnography of a Category*. Duke University Press.

Weinreich, U.; Labov, W.; and Herzog, M. I. 1968. Empirical Foundations for a Theory of Language Change. In *Directions for Historical Linguistics*, 95–188. Austin, TX: University of Texas Press.

Zimman, L. 2014. The Discursive Construction of Sex: Remaking and Reclaiming the Gendered Body in Talk about Genitals among Trans Men. In *Queer Excursions: Retheorizing Binaries in Language, Gender, and Sexuality*, 13–34. Oxford: Oxford University Press.

Zimman, L.; and Brown, C. 2022. "Critics," "boosters" and the politics of linguistic change: A computational analysis of the lexicon in an online trans community. Paper presented at the Linguistic Society of America Conference, D.C., USA.

Zimman, L.; and Hayworth, W. 2020. How we Got Here: Short-Scale Change in Identity Labels for Trans, Cis, and Non-Binary People in the 2000s. In *Proceedings of the Linguistic Society of America*, volume 5, 499–513.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes; we discuss ethical concerns in the ethics statement above.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes; we discuss the integration of quantitative and qualitative analyses, as well as the motivations for the various quantitative analyses we conduct.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, in the limitations section.
   (e) Did you describe the limitations of your work? Yes, in the limitations section.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, in the ethics statement above.
   (g) Did you discuss any potential misuse of your work? Yes, in the ethics statement above.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes; we discuss anonymization of examples in the ethics statement above.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing... Our study is exploratory in nature and does not involve formal hypothesis testing, but we have addressed some of these concerns.
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes; we have stated the assumptions of key methods
   (b) Have you provided justifications for all theoretical results? Yes; we have justified our interpretations
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes; in our results sections (Sections 5.3 and 6.2 to 6.4), we discuss alternative interpretations
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes; in our results sections (Sections 5.3 and 6.2 to 6.4), we discuss alternative interpretations
   (e) Did you address potential biases or limitations in your theoretical framework? Yes, in the limitations section
   (f) Have you related your theoretical results to the existing literature in social science? Yes, throughout.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, in Section 7
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments... We do not consider the snapshot language model analysis to be what is referred to by "machine learning experiments" here, but we have indicated answers that are relevant for our study
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No, because the data are not publicly available, as stated in the ethics statement above
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes; we have given all details of the snapshot language models
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, all figures contain 95% confidence intervals based on resampling
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA; we have not evaluated model performance in this sense
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes, we mentioned that the data is not publicly available in the ethics statement above
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, in the ethics statement above
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, in the ethics statement above
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA

Ethics Statement

We obtained our data from the creators of the TransLiveCorpus (Zimman and Hayworth 2020) directly; the corpus is not widely available and has no distribution license. The LiveJournal message board is publicly accessible, and the corpus was compiled through automated web-scraping, without obtaining individualized user consent. The metadata does not contain identifiable information, but users do discuss highly personal and sometimes offensive content, and there are ethical questions raised by analyzing words posted by a marginalized community members. Consequently, we anonymized entries that we presented as examples here by slightly rephrasing them to limit searchability and by using pseudo-usernames and pseudo-avatars, a practice done by other studies of online forums (e.g., Thach et al. 2024).

In this current moment of high politicization of trans topics, there is a risk that transphobic critics could use research insights into trans communities against them. We see this potential risk as threefold: firstly that individuals could be targeted based on this paper; secondly that the argument may be stripped of nuance and co-opted by transphobic critics; and thirdly, that our discussion of topics could be using framing or terminology that characterizes trans people in unintentionally oppressive ways. We address each of these points in line with the Belmont Report's principle of "do no harm". Regarding the first point, this paper focuses on a now-disused community, rephrases example quotes for limited searchability, and does not reveal private insider information that could put individuals at risk. Regarding the second, we acknowledge that part of our argument (that community-internal oppressive dynamics within trans spaces may come into play in the development of community language norms) could be taken out of context and appropriated to align with discourse calling for lack of protection for trans people online. However, we believe that it is crucially important for those in marginalized communities to be cognizant of the ways that other power structures and dynamics show up within communities. In seeing the fight against transphobia as not isolated from other struggles for justice and liberation, it is important to understand ways that community discourse can exclude and divide. In our language and framing within the paper, we strove to be sensitive with respect to this tension. Regarding the third point, as the research team includes two transmasc researchers with experience participating in trans-specific online forums, one of whom participated in the community studied, community-insider knowledge was drawn upon in framing content in a way sensitive to community concerns.

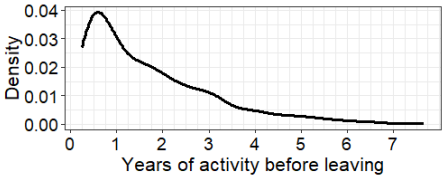

Figure 10: Density plot of the number of years that leavers in our analysis were active in the community prior to leaving.

A Activity Over Time

Appendix A shows the number of users and entries in each time bin that featured in our analysis (after applying the filters described in Section 5.1), broken down by user group.

In the period we analyze, the number of active 'everyday' users in the community (as reflected by the remainers) started growing in 2004, peaked in 2007, and declined back to early levels by 2009. Associated with this rise and fall in activity was a turnover of population: while there was a steady stream of newcomers and leavers throughout, large numbers of new members joined the community leading into and through the period of peak activity, and increasing numbers of members left the community during and after the period of peak activity. There was a similar population turnover among the prolific users: some of the new members that joined during the period of growth became extremely active, increasing the number of prolifics, while some of the extremely active members were among those to leave after the activity peak, reducing the number of prolifics.

Population dynamics aside, it is clear that the four groups are well differentiated by their activity levels. Prolifics were consistently more active than the 'everyday' remainers, and levels of activity were less skewed among prolifics than among remainers. Members were less active when they joined the community as newcomers than when they became entrenched remainers, and even less active in the lead-up to their leaving the community as leavers.

Figure 10 shows the distribution over the length of time that leavers in our analysis were active in the community prior to leaving. Due to our filtering criteria, all leavers included in the analysis were active for at least 90 days. While there is a peak in users leaving after 6–9 months of activity, most users remain active in the community for more than a year before leaving, and many remain active for multiple years; the mean activity duration for leavers is 20.4 months, and the median is 16 months.

B The Choice of Time Period

Our analysis spanned from mid-2003 to mid-2009: a subset of the time period represented in the corpus. Our choice of this period was motivated by the fact that it contains the highest activity in the community (as demonstrated in Appendix A) and by considerations of training data size.

Time bin		Remainers			Prolifics			Newcomers			Leavers		
		Users	Entries/User		Users	Entries/User		Users	Entries/User		Users	Entries/User	
year	*months*		*mean*	*med.*		*mean*	*med.*		*mean*	*med.*		*mean*	*med.*
2003	Jul–Sep	238	9.2	5	6	51.2	49.5	67	6.6	4	28	3.3	1
2003	Oct–Dec	233	8.5	5	6	42.8	43.5	50	6.4	4	28	4.8	2
2004	Jan–Mar	236	8.0	5	7	47.0	42	72	7.7	4	42	5.1	2
2004	Apr–Jun	266	8.7	5	8	49.4	56.5	79	6.2	3	30	4.0	2
2004	Jul–Sep	271	9.2	5	8	43.1	51	74	7.0	3	30	3.9	2
2004	Oct–Dec	292	9.1	5	11	32.7	24	93	6.3	2	45	3.9	2
2005	Jan–Mar	369	11.5	5	11	53.3	59	93	13.3	3	49	3.3	2
2005	Apr–Jun	355	9.1	4	11	33.7	31	83	5.7	2	49	3.7	2
2005	Jul–Sep	375	9.8	5	10	57.4	55.5	93	7.0	3	47	5.4	2
2005	Oct–Dec	384	8.1	4	10	51.2	48.5	67	7.5	4	44	4.1	1.5
2006	Jan–Mar	394	9.4	4.5	11	46.0	44	99	6.4	4	59	5.4	2
2006	Apr–Jun	417	10.4	5	10	67.9	61.5	97	8.2	4	70	3.9	2
2006	Jul–Sep	443	12.1	6	12	73.8	73	101	6.1	4	64	5.0	2
2006	Oct–Dec	463	10.1	5	12	77.2	67.5	89	8.3	4	56	4.5	2
2007	Jan–Mar	475	11.5	5	11	99.2	101	95	9.1	4	81	3.9	2
2007	Apr–Jun	467	10.5	5	12	56.2	43	94	7.1	4	91	4.2	2
2007	Jul–Sep	458	8.6	4	11	59.3	58	84	6.2	3	82	3.7	2
2007	Oct–Dec	467	8.3	4	12	51.8	51	87	10.8	5	79	3.3	2
2008	Jan–Mar	445	7.8	4	10	41.5	40.5	65	6.4	3	80	3.2	1
2008	Apr–Jun	436	6.1	3	10	25.7	24.5	34	4.6	3	72	2.3	1
2008	Jul–Sep	383	5.7	3	10	27.5	19	34	5.1	3	75	2.4	1
2008	Oct–Dec	353	5.6	3	10	22.7	16.5	35	3.1	2	67	2.3	1
2009	Jan–Mar	348	5.4	3	9	24.6	24	35	3.2	2	67	3.0	2
2009	Apr–Jun	295	4.5	3	9	16.6	12	17	3.5	1	49	2.1	1

Table 2: Descriptive statistics for each user group within each 3-month time bin analyzed: number of users posting in the bin, and mean and median number of entries posted per user in the bin.

Appendix B summarizes the descriptive statistics that we considered for each 3-month time bin in order to determine whether it would meet our requirements for training data size. As described in Section 5.2, our analysis required us to resample training sets of 500 entries (containing at least 30 tokens each) from each time bin, composed of 3 entries from each of 113 users, 2 entries from each of 53 different users, and 1 entry from each of 55 yet different users. To ensure that resampling would capture sufficient variation in potential training sets, we only included time bins in which these minimum criteria could be exceeded by at least 15%. That is, we only included time bins in which there were at least 130 remainers posting 3+ times, 191 remainers posting 2 or 3+ times, and 255 remainers posting 1, 2, or 3+ times, supporting the formation of a training set of at least 575 entries without taking more than 3 entries from any single user. The only time bins that met this criteria were between mid-2003 and mid-2009; while there were many entries posted outside of this period, no 3-month time bin had enough entries from enough distinct users to support the formation of adequately controlled training sets.

We also considered time bins that were smaller or larger than 3 months. Smaller time bins (1 or 2 months) gave greater temporal granularity, but came with the cost of further limiting the period of analysis because they yielded fewer entries (from fewer users) in each time bin. Larger time bins (4 or 6 months) reduced the temporal granularity and did not allow the period of analysis to be extended more than a few months. Our choice of 3-month time bins therefore reflects an optimal balance of high temporal granularity and a long period of analysis.

C *bio-/cis-* Terms

Appendix C contains the lemmas that are modified by *bio-* and/or *cis-* in relation to gender identity multiple times in the corpus. Lemmas included in the previous analysis by Zimman and Hayworth (2020) are underlined.

The following lemmas are modified by *bio-* in relation to identity (as an adjective or prefix) once in the corpus: *adult*, *anything*, *bit*, *bottom*, *chest*, *couple*, *example*, *genitals*, *mother*, *organ*, *original*, *pronoun*, *puberty*, *queen*, *reference*, *scrotum*, *sexually*, *sibling*, *sister*, *sized*, *son*, *teen*, *testicle*, *testicles*, *twin*, *variety*.

The following lemmas are modified by *cis-* in relation to identity (as an adjective or prefix) once in the corpus: *acquaintance*, *aunt*, *binary*, *chick*, *community*, *fella*, *forum*, *genderness*, GLB, *grandmother*, *kid*, *lady*, *lesbian*, *level*, *media*, *member*, *mom*, *norm*, *normative*, *patient*, *public*, *researcher*, *roommate*, *society*, *standard*, *straight*, *student*, *wannabe*, *whathaveyou*, *word*.

D Calculating Entropy and Divergence

The decomposition of cross-entropy into entropy and divergence is a novel extension of the framework of snapshot language models introduced by Danescu-Niculescu-Mizil et al. (2013). Using this decomposition to compare across snapshot language models and test sets is only made appropriate by our efforts to control the training and test sets, because the noise level in maximum likelihood estimation of a language model is (inversely) correlated with the size of the training set and entropy is correlated with the size of the test set (via the number of unique types that it contains). The technical details of the decomposition are as follows.

For a given test set $T_{i,j}^{(b)}$ of 100 entries $e_k^{(b)}$ from time bin b, we extracted the first, middle, and last 10-token snippets from each entry, from which we formed bigram[12] units $u_{k,1}^{(b)}, \ldots, u_{k,30}^{(b)}$. Given a snapshot language model $\text{SLM}_i^{(b)}$ from time bin b, trained on similar 30-token snippets from 500 entries, we calculated the *per-unit cross-entropy* of each entry in the test set with respect to the model, $H(e_k^{(b)}, \text{SLM}_i^{(b)})$:

$$H(e_k^{(b)}, \text{SLM}_i^{(b)}) = \frac{-1}{|e_k^{(b)}|} \sum_{u_{k,n}^{(b)} \in e_k^{(b)}} \log_2 P_{\text{SLM}_i^{(b)}}(u_{k,n}^{(b)}) \quad (1)$$

where $P_{\text{SLM}_i^{(b)}}(u)$ is the (joint) probability the model assigns to unit u. The higher an entry's per-unit cross-entropy, the less adequately the model can statistically predict it. High cross-entropy indicates that the entry has low consistency with the linguistic conventions inferred from the 500 entries that the model was trained on.

To characterize how well a snapshot language model predicts language use in test set entries *in general*, we then calculated the *average* per-unit cross-entropy over all entries in the test set, $H(T_{i,j}^{(b)}, \text{SLM}_i^{(b)})$:

$$H(T_{i,j}^{(b)}, \text{SLM}_i^{(b)}) = \frac{1}{|T_{i,j}^{(b)}|} \sum_{e_k^{(b)} \in T_{i,j}^{(b)}} H(e_k^{(b)}, \text{SLM}_i^{(b)}) \quad (2)$$

The higher the average per-unit cross-entropy of a snapshot language model, the less adequately the model statistically predicts language use in the test set as a whole. This indicates that any linguistic conventions inferred from the 500 entries that the model was trained on do not robustly generalize to the 100 other contemporaneous entries in the test set, implying linguistic variation.

When applying a model $\text{SLM}_i^{(b)}$, the test entries $T_{i,j}^{(b)}$ are coerced to the inventory of units that it supports, by replacing tokens that do not occur in the corresponding training entries with a common *UNK* token. Thus, average per-unit cross-entropy can be conceptualized as the cross-entropy between the empirical probability distribution over units in the test entries, $P_{T_{i,j}^{(b)}}(u)$, and the modeled distribution over units in the training entries, $P_{\text{SLM}_i^{(b)}}(u)$:

$$H(T_{i,j}^{(b)}, \text{SLM}_i^{(b)}) = - \sum_{u \in e_k^{(b)} \in T_{i,j}^{(b)}} P_{T_{i,j}^{(b)}}(u) \log_2 P_{\text{SLM}_i^{(b)}}(u) \quad (3)$$

where $P_{T_{i,j}^{(b)}}(u)$ represents the relative frequency of each unit u across all 30-unit snippets in $T_{i,j}^{(b)}$:

Time bin			Remainers posting...			
year	*months*	*entries*	*1 time*	*2 times*	*3+ times*	*max. train*
2001	Jan–Mar	28	1	0	5	16
2001	Apr–Jun	27	4	2	3	17
2001	Jul–Sep	39	7	5	6	35
2001	Oct–Dec	18	6	3	2	18
2002	Jan–Mar	238	5	7	27	100
2002	Apr–Jun	997	11	7	62	211
2002	Jul–Sep	1284	19	12	88	307
2002	Oct–Dec	1572	18	18	101	357
2003	Jan–Mar	1976	29	16	146	499
2003	Apr–Jun	1918	42	22	153	545
2003	Jul–Sep	2192	40	26	172	608
2003	Oct–Dec	1988	44	30	159	581
2004	Jan–Mar	1884	41	28	167	598
2004	Apr–Jun	2316	48	33	185	669
2004	Jul–Sep	2493	39	40	192	695
2004	Oct–Dec	2660	67	36	189	706
2005	Jan–Mar	4252	67	42	260	931
2005	Apr–Jun	3238	82	39	234	862
2005	Jul–Sep	3659	77	46	252	925
2005	Oct–Dec	3117	82	41	261	947
2006	Jan–Mar	3688	72	49	273	989
2006	Apr–Jun	4336	76	50	291	1049
2006	Jul–Sep	5373	75	50	318	1129
2006	Oct–Dec	4669	92	60	311	1145
2007	Jan–Mar	5452	90	57	328	1188
2007	Apr–Jun	4901	75	61	331	1190
2007	Jul–Sep	3952	79	57	322	1159
2007	Oct–Dec	3868	110	72	285	1109
2008	Jan–Mar	3458	88	60	297	1099
2008	Apr–Jun	2665	116	64	256	1012
2008	Jul–Sep	2197	100	64	219	885
2008	Oct–Dec	1981	91	62	200	815
2009	Jan–Mar	1895	96	55	197	797
2009	Apr–Jun	1315	94	50	151	647
2009	Jul–Sep	972	89	54	116	545
2009	Oct–Dec	968	77	48	103	482
2010	Jan–Mar	1240	70	41	121	515
2010	Apr–Jun	998	91	43	103	486
2010	Jul–Sep	1142	72	45	122	528
2010	Oct–Dec	820	75	42	100	459
2011	Jan–Mar	872	75	34	105	458
2011	Apr–Jun	685	70	32	77	365
2011	Jul–Sep	504	47	28	68	307
2011	Oct–Dec	660	54	31	69	323
2012	Jan–Mar	395	55	34	49	270
2012	Apr–Jun	556	37	33	67	304
2012	Jul–Sep	326	37	25	37	198
2012	Oct–Dec	173	33	16	24	137

Table 3: Descriptive statistics used to determine 3-month time bins with adequate training data: total number of entries posted by remainers that contain at least 30 tokens; number of remainers posting once, twice, or 3+ times during the time bin; and maximum number of entries in the training set based on not taking more than 3 entries from any single remainer. The middle portion of the table separated by horizontal lines, from mid-2003 through mid-2009, contains all of the time bins that have enough entries from enough distinct users to support the formation of adequately controlled training sets.

Lemma	*bio-* adj.	*bio-* prefix	*cis-* adj.	*cis-* prefix	**Total**
guy	462	537	90	69	**1158**
male	254	371	53	64	**742**
man	216	198	150	65	**629**
gendered	1	0	2	549	**552**
gender	5	3	4	479	**491**
woman	36	39	82	56	**213**
female	55	94	16	6	**171**
boy	85	75	2	4	**166**
person	6	1	103	14	**124**
sexual	0	1	0	73	**74**
girl	20	40	5	7	**72**
penis	21	11	0	0	**32**
dick	13	13	1	0	**27**
sexist	0	0	1	26	**27**
friend	11	0	13	1	**25**
cock	5	6	9	1	**21**
dude	5	5	3	1	**14**
folk	0	0	12	2	**14**
brother	2	8	2	0	**12**
name	9	2	0	0	**11**
fag	3	6	0	0	**9**
king	9	0	0	0	**9**
sexism	0	0	0	9	**9**
genderism	0	0	0	8	**8**
boyfriend	3	0	2	2	**7**
privilege	0	0	3	4	**7**
centric	0	4	0	1	**5**
sex	4	0	0	1	**5**
status	2	1	2	0	**5**
born	3	1	0	0	**4**
whatever	0	4	0	0	**4**
partner	0	0	4	0	**4**
scum	0	0	4	0	**4**
sexed	0	0	0	4	**4**
centrism	0	3	0	0	**3**
father	3	0	0	0	**3**
girlfriend	2	0	1	0	**3**
ally	0	0	3	0	**3**
cissy	0	0	0	3	**3**
individual	0	0	3	0	**3**
centrist	0	2	0	0	**2**
femme	0	2	0	0	**2**
junk	1	1	0	0	**2**
papa	0	2	0	0	**2**
way	2	0	0	0	**2**
body	1	0	1	0	**2**
dad	1	0	1	0	**2**
term	1	0	1	0	**2**
uncle	1	0	1	0	**2**
world	1	0	1	0	**2**
gay	0	0	0	2	**2**
genderistic	0	0	0	2	**2**

Table 4: Lemmas that are modified by *bio-* and/or *cis-* in relation to gender identity multiple times in the corpus.

$$P_{T_{i,j}^{(b)}}(u) = \frac{\sum_{e_k^{(b)} \in T_{i,j}^{(b)}} \text{count}(u \text{ in } e_k^{(b)})}{\sum_{e_k^{(b)} \in T_{i,j}^{(b)}} |e_k^{(b)}|} \quad (4)$$

Leveraging this conceptualization, we decomposed cross-entropy into *entropy*, $H(T_{i,j}^{(b)})$, and *divergence*, $D_{\text{KL}}(T_{i,j}^{(b)} \| \text{SLM}_i^{(b)})$. We calculated entropy from the empirical probability distribution over units (from the training inventory) in the test entries, and we calculated divergence by subtracting entropy from average per-unit cross-entropy:

$$H(T_{i,j}^{(b)}) = -\sum_{u \in T_{i,j}^{(b)}} P_{T_{i,j}^{(b)}}(u) \log_2 P_{T_{i,j}^{(b)}}(u) \quad (5)$$

$$D_{\text{KL}}(T_{i,j}^{(b)} \| \text{SLM}_i^{(b)}) = H(T_{i,j}^{(b)}, \text{SLM}_i^{(b)}) - H(T_{i,j}^{(b)}) \quad (6)$$

This decomposition reflects the idea that there are two reasons why linguistic conventions inferred from entries in the training set do not robustly generalize to entries in the test set: either entries in the test set may not consistently follow clear linguistic conventions at all (high entropy), or they may follow conventions that are different to those observed in entries in the training set (high divergence).

For each snapshot language model $\text{SLM}_i^{(b)}$, we averaged the entropy and divergence values across the 100 sampled test sets $T_{i,1}^{(b)}, \ldots, T_{i,100}^{(b)}$ to get single point-estimates of the expected entropy and divergence under the corresponding choice of training entries. We then formed distributions of these point-estimates over the 100 snapshot language models $\text{SLM}_1^{(b)}, \ldots, \text{SLM}_{100}^{(b)}$ in each time bin b, and we tracked the means of these distributions across time bins. This gave us estimated trajectories for entropy and divergence that are independent of specific choices of test and training data.