Explain like I am a Scientist: The Linguistic Barriers of Entry to r/science

Tal August[1], **Dallas Card**[2], **Gary Hsieh**[3], **Noah A. Smith**[1,4], **Katharina Reinecke**[1]
[1]Paul G. Allen School of CSE, University of Washington, [2]Stanford University, [3]Human Centered Design and Engineering, University of Washington, [4]Allen Institute for Artificial Intelligence
{taugust, nasmith, reinecke}@cs.uw.edu, dcard@stanford.edu, garyhs@uw.edu

ABSTRACT
As an online community for discussing research findings, *r/science* has the potential to contribute to science outreach and communication with a broad audience. Yet previous work suggests that most of the active contributors on *r/science* are science-educated people rather than a lay general public. One potential reason is that *r/science* contributors might use a different, more specialized language than used in other subreddits. To investigate this possibility, we analyzed the language used in more than 68 million posts and comments from 12 subreddits from 2018. We show that *r/science* uses a specialized language that is distinct from other subreddits. Transient (newer) authors of posts and comments on *r/science* use less specialized language than more frequent authors, and those that leave the community use less specialized language than those that stay, even when comparing their first comments. These findings suggest that the specialized language used in *r/science* has a gatekeeping effect, preventing participation by people whose language does not align with that used in *r/science*. By characterizing *r/science*'s specialized language, we contribute guidelines and tools for increasing the number of contributors in *r/science*.

Author Keywords
Science communication, online communities, language style

CCS Concepts
•**Human-centered computing** → **Empirical studies in collaborative and social computing;** *Empirical studies in HCI;*

INTRODUCTION
Science communication is important for both scientists and the public as it allows communicating and discussing research findings with a broad audience [36]. While scientific findings have traditionally been curated by journalists, science communication has become more scalable and democratized with the advent of the Internet, which has enabled sharing of scientific findings in blog posts, social networks, or other online communities.

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376524

r/science, a sub forum ("subreddit") on the social news aggregation platform Reddit and shown in Figure 1, has emerged as one of the largest platforms for disseminating and discussing scientific findings outside of the social circles of the scientists themselves. However, past work has shown that active contributors in *r/science* are largely already involved in scientific activity, and broader public engagement is lacking [17].

A possible reason why *r/science* does not attract broader contribution is that the community might have developed specific norms that deter some users from actively participating. Clashes between user values and community norms, or users failing to adopt community norms, can lead to negative community feedback and reduced engagement [10, 28]. One such norm is the specialized language members might use to contribute to an online community [8]. Commonly used insider words [8], politeness or formality [2, 27], and even pronoun usage [35] are all examples of linguistic behaviors developed within a community that characterize its language. Just as with other normative behaviors, the language is important for new users to adopt: users who don't adopt it receive fewer and less supportive responses [16, 32] and are more likely to leave an online community [8]. Given that science communication is often hindered by specialized language (e.g., jargon) [29], it is possible that the *r/science* community has developed language norms that prevent a diverse public from engaging.

To explore whether *r/science* contains specialized language that may present a barrier to entry, we analyzed differences in the language of 68,560,317 publicly available posts and comments on Reddit. We compared the language used in *r/science* to the language used by contributors of 11 other subreddits using language models, a commonly used technique for measuring language differences [8].

We found our hypothesis to be true: the language used in *r/science* posts and comments differs from the language used in other large or topically related subreddits. Many of the distinctions in language in *r/science* reflect *r/science*'s rules for language norms, such as hedging (qualifying statements with words such as 'suggest' and 'possibly'), impersonal language, and scientific terminology. Our results show that transient contributors (i.e., those who only post once) on *r/science* fail to adapt to this specialized language more often than more experienced authors, suggesting that *r/science*'s language norms do not always come naturally to people.

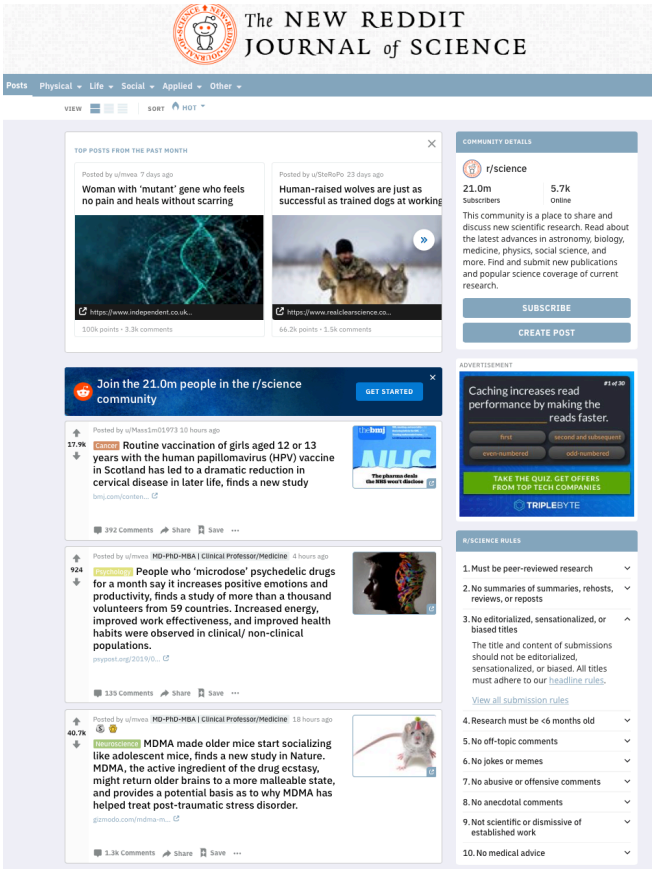

Figure 1. The main page of *r/science* with top posts on the left and a description of the community and its rules on the right.

Our findings indicate that *r/science*'s guidelines and community norms, while useful to maintain a high standard of rigor and discourse, have the side effect of limiting contributions from a broader audience by enforcing a specialized language that people wanting to post and comment first have to learn. While our analysis was on *r/science*, our methodology can help researchers identify if there are similar gatekeeping effects of specialized language in other scientific communication, such as blogs and social media platforms. Our work also provides several design implications, such as how to guide newcomers in contributing and following language norms, and ways to relax some rules without sacrificing scientific rigor.

RELATED WORK

Related to the current paper are (1) research on science communication, (2) work on online community norms and guidelines, and (3) studies of specialized language in online communities.

Science Communication

Science communication is the use of appropriate skills, media, activities, and dialogue to produce one or more of the following personal responses to science: awareness, enjoyment, interest, opinions, and understanding [3]. Over the past twenty years, the focus of science communication has shifted from dissemination to dialog and participation [14]. The goal is no longer simply to provide the public with more information to overcome any information deficits, but to increase public engagement through two-way interactions.

Advances in social media and social technologies are offering novel and hybrid forums to support interactions between scientists and different publics. Not only are people turning to blogs and online-only media sources for science information, the social and interactive features of Web 2.0 are also allowing people to produce and discuss science online [1]. One such popular example is *r/science*, a subreddit on Reddit. A study of *r/science*'s 2016 posts suggests that it is a vibrant community attracting a variety of science information and discussions [17]. The most frequent comments posted include: questions about the research or related work; extending, applying, or reasoning about the research; personal questions or stories or responses to such; and offering an educational response.

However, one concern identified in the study of *r/science* is that despite its potential for supporting dialogue, it may only attract and sustain participation from an interested and knowledgeable public [17]. A noted tension was between facilitating broad science dialogues and ensuring high quality information. *r/science* ensures high quality information with a list of relatively strict guidelines on their front page (see bottom right corner in Figure 1), such as requiring all posts to link to peer-reviewed research and disallowing sensationalized post headlines. This is especially evident when an *r/science* post becomes popular enough to reach *r/all* (a default page showing a digest of the most popular posts from a user's subscriptions). During these events, Reddit users who are not familiar with *r/science* norms can become exposed to *r/science* content. However, when these new *r/science* visitors post comments, their comments are often not aligned with *r/science*'s norms, resulting in many of the comments getting deleted[17]. This could be discouraging for newcomers and the lay public and inhibit them from further participation – a hypothesis that we seek to investigate in this paper.

Online Community Norms and Guidelines

Online communities develop norms and guidelines that users follow to be productive members of a community. These guidelines are often decided as a rough consensus of members around what behavior is or is not acceptable in the community [18]. There are some community standards governing the entire Reddit platform ("reddiquette"), but it is common for individual subreddits to have their own set of specialized rules and norms [9]. Chandrasekharan et al. [5] characterized Reddit community rules into three levels: Macro (i.e., rules shared across all of Reddit), Meso (i.e., rules shared across groups of subreddits), and Micro (i.e., subreddit specific rules). More popular subreddits usually have more structured norms due to having to handle and socialize large influxes of new members without losing community values [9]. Moderators also play a strong role in developing and enforcing community norms, which can range from being highly active in a community, such as posting content often, to only being involved when a member seriously violates a rule [31].

Although community norms are often publicly displayed, newcomers can be overwhelmed by these rules, risking community rejection by unknowingly violating norms [12, 18], or turned

off from the community by guidelines clashing with their own values [28]. Research suggests that Wikipedia's sharp decline in retention of desirable new editors (i.e., not vandals) from around 40% in 2003 to less than 10% in 2010 can partly be attributed to inflexible rules [11]. Community guidelines can also deter some users from joining a community due to an unintended clash in values, such as StackOverflow's rule of "No thank you's", that conflicts with many users' beliefs around healthy community support [28].

Publicly displaying community norms supports new users in interacting with the norms of the community, which can help increase normative behavior in newcomers [22]. Cialdini et al. [7] characterized two ways norms influence behavior: injunctively, where norms prescribe acceptable behaviours in the group, and descriptively, where member behaviour provides examples of norms. Morgan and Filippova [24] characterized these injunctive and descriptive norms in Wikipedia subcommunities, identifying injunctive norms as posted guidelines and descriptive norms as active community threads. Both injunctive and descriptive norms can increase normative behaviour in online communities, especially when injunctive norms are reinforced with descriptive norms, or vice versa.

While many of the guidelines and moderation in online communities focus on acceptable behaviors, like how people should treat other members or what they can post about, communities also develop unique *language* norms, such as specific words members use [8] that are important for new members to follow in order to receive support from the community [8, 16, 32]. In this paper, we extend prior work by exploring the language norms that *r/science* has developed and showing how these norms can act as an additional gatekeeper.

Community Language

Studying community language norms has a long history in sociolinguistic research. Labov [20] studied fine-grained phonetic differences in New York City, showing that different socio-economic classes, and even small peer groups within these classes, use significantly different vocalizations, such as a different pronunciation of the /r/ sound [19, 20]. Milroy and Milroy [23] showed in their seminal work exploring vernacular English in Belfast, Northern Ireland, that community networks, such as familial ties, were a strong influence on an individual's linguistic variation.

Research has drawn comparable findings for linguistic variation in online communities [8, 35, 27]. Cassell and Tversky [4] explored the formation of a new online community comprised of children from around the world, finding that these children from diverse cultural, economic, and geographic backgrounds converged on a shared language style, such as speaking in the collective voice, and topics of conversation. Danescu-Niculescu-Mizil et al. [8] had similar findings in online beer enthusiast communities, showing that members adopt certain words (e.g., "aroma" or fruit-related words) that become widespread in the community. Tran and Ostendorf [35] characterized the language style of 8 subreddit communities, showing that stylistic features, such as community-specific jargon and sentence structure, led to close to 90% accuracy in identifying a community. Zhang et al. [37] built on this research, exploring a larger subset of 300 subreddits and showing that frequently used words within a subreddit were also useful in characterizing how the community was distinctive (different from other communities) and dynamic (how quickly the community shifted to new topics). They found that distinctive and dynamic communities are more likely to retain users.

Table 1. Number of posts, comments, and subscribers of each subreddit.

Subreddit	# Posts	# Comments	# Subscribers
r/science	22,157	604,267	21,015,665
r/news	254,201	5,878,711	17,972,696
r/politics	211,866	14,754,150	4,925,536
r/pics	101,624	3,806,068	21,313,781
r/funny	114,272	4,246,111	23,759,930
r/askreddit	1,096,947	35,665,797	22,153,598
r/askhistorians	41,203	72,056	998,325
r/everythingscience	6,772	36,901	165,852
r/futurology	20,127	799,176	14,037,974
r/truereddit	4,193	198,108	439,334
r/dataisbeautiful	8,437	420,762	13,608,622
r/askscience	12,162	184,249	17,901,914

Adopting the language style of a community is important for being accepted by the community [2, 8, 16]. For example, users who did not adopt the specific words common in the beer enthusiast communities mentioned above were more likely to leave compared to members who did adopt these new words [8]. Members of breast cancer online support groups use much more community-specific jargon and informal language the longer they remain part of the community, indicating that language style is a reflection of community socialization [27]. In addition, members of mental health support groups on Reddit are more likely to receive supportive responses if their language matches the style of the community [32], and Twitter users who match the language style of their followers receive more retweets [34]. We build on this work by exploring the language norms of *r/science*, and how new members in *r/science* adopt, or don't adopt, these norms.

METHOD

We conducted linguistic and quantitative analyses of 12 large subreddits to answer our overarching question of whether *r/science*'s potentially specialized language represents a barrier of entry. More precisely, our research questions are:

RQ1: Do posts and comments on *r/science* contain specialized language compared to other large or topically related subreddits? If so, what are the characteristics of this specialized language?

RQ2: How does the language of users who stay and leave differ in their posts and comments? If there is a defined difference, this would suggest that new contributors experience a language barrier, and those who join *r/science* with a strongly differing language are less likely to stay than those whose language more closely aligns with *r/science*'s.

Data

In addition to *r/science*, we selected 11 subreddits with the goal of comparing the language used in *r/science* to

the language used in other large or topically related communities: *r/news, r/politics, r/pics, r/funny, r/askhistorians, r/futurology, r/truereddit, r/dataisbeautiful, r/askscience, r/everythingscience* and *r/askreddit*. While there is no such thing as "the average Reddit user" that we could compare against, our list includes some of the largest subreddits that share a similar subscriber count to *r/science* (*r/news, r/politics, r/pics, r/funny*, and *r/askreddit*) and those with the largest overlap in user participation with *r/science* (i.e., users commenting and posting in *r/science* and the other subreddit): *r/askhistorians, r/futurology, r/truereddit, r/dataisbeautiful, r/askscience*, and *r/everythingscience*) [21]. All of these are subreddits with community members that *r/science* should have an interest in engaging, which is why our aim was to characterize how different the language experience is for a user in these subreddits compared to *r/science*.

For each subreddit, we collected all comments and posts from Jan. 2018 to Dec. 2018 using Google's Bigquery.[1] We ignored comments and posts that had been deleted by their original author, by moderators, or that were shorter than 10 words, for a total of 1,893,961 posts and 66,666,356 comments. Table 1 provides details of our dataset.

Analysis

RQ1: *r/science*'s specialized language

We analyzed whether *r/science* uses specialized language by constructing language models trained on its posts and comments. A language model is a conditional probability distribution over each word in a vocabulary given preceding words (left context); the distributions are estimated using a training corpus of texts, which we denote $\mathscr{D}$. A language model can be used to assign probability mass to a sequence of words by taking the product of conditional probabilities for individual words given their left contexts (i.e., applying the chain rule of probability). The probability that a language model trained on text dataset $\mathscr{D}$, which we denote $LM(\mathscr{D})$, assigns to a new piece of text depends very heavily on the training data $\mathscr{D}$. For example, training a language model on judicial decisions will likely lead to very low probability assignment for a Reddit post about astronomy, but higher probability for a similar-length legal brief.

Given a language model $LM(\mathscr{D})$, we can calculate how different a text (word sequence $\vec{w} = \langle w_1, w_2, \ldots, w_N \rangle$) is from the training data of LM by calculating $\vec{w}$'s **cross entropy** under LM:

$$CE(\vec{w}, LM(\mathscr{D})) = -\frac{1}{N}\sum_{i=1}^{N} \log p_{LM(\mathscr{D})}(w_i \mid w_{i-1}), \quad (1)$$

where $p_{LM(\mathscr{D})}(w_i \mid w_{i-1})$ is the "bigram" probability of word w_i given the preceding word w_{i-1} according to the language model $LM(\mathscr{D})$, and w_0 and w_N are special "start" and "stop" tokens included by convention.[2] The higher the cross entropy, the more divergent $\vec{w}$ is from $LM(\mathscr{D})$'s training data, the less likely it is under $p_{LM(\mathscr{D})}$, and hence the more "surprising" it is under the distribution of language model $LM(\mathscr{D})$.

Our language models are Katz-backoff bigram models with Good-Turing smoothing [6]—a commonly used technique to improve probability estimates that past work in this space has employed [37]—trained on posts and comments of active authors in *r/science*. All language models used as a vocabulary the words seen during training. We built our language models using the SRILM toolkit [33].

For comments, our language models were trained on 1000 comments, 5 comments sampled from 200 experienced commenters for each month. We defined experienced commenters as those who had commented at least 5 times in a given month following [37]. For posts, the language models were trained on 1000 posts (5 posts sampled from 200 experienced posters for the entire year). We treated experienced posters as those who had posted at least 5 times in the year. This provided roughly the same percentage of users per year that our definition of experienced commenters did for a month. We constructed 100 language models for each month of comments and 100 language models total for the year's posts, resampling authors and their comments and posts each time. We used all language models in each cross entropy calculation for a post or comment, averaging all post or comment cross entropy within a language model.

Past work has identified that longer posts and comments tend to exhibit higher cross entropy, possibly due to the higher probability of esoteric language in longer posts or comments [8, 37]. To avoid these length effects, we followed past work [37] and only used the first ten words of each comment or post for training and cross entropy calculations, inserting stop tokens at the end of these first ten words.[3]

With language models trained on *r/science* posts and comments, we analyzed whether text from other subreddits diverged from the text of *r/science* by calculating the cross entropy of text sampled from other subreddits and comparing it to text sampled from *r/science*. In particular, we calculated the cross entropy of 50 comments, 5 unique comments sampled from 10 experienced authors (different from those used to train the models), for each month from each subreddit using language models trained on *r/science* comments for that month. We then conducted an ANOVA and Tukey post-hoc tests to compare the cross entropy of *r/science* comments compared to the comments of other subreddits, averaged over all months. We conducted the same analysis for *r/science* posts, sampling over the entire year 12 times to match the number of month samples.

One limitation of using language models trained on *r/science* is that they will overfit to *r/science* posts and comments. This leads us to expect that they will find posts and comments of other subreddits surprising (i.e., have higher cross entropy). All a higher cross entropy means is that something in the text of *r/science* is different from other subreddits, but it doesn't

[1] https://cloud.google.com/bigquery/

[2] While bigram models are a relatively simple choice, they suffice for our purposes and are relatively robust against overfitting for datasets of the size we consider here.

[3] We obtained qualitatively similar results using the entire span of each post or comment for training.

signal what that something is. To characterize the actual differences between the language used in *r/science*, we therefore additionally analyzed individual word frequencies using uni- and bigram language models for both posts and comments (separately). To do this we first estimated an empirical background frequency for each token (counted as a unigram or bigram) using all other subreddits:

$$\hat{p}_w = (c_w + \alpha) \,/ \sum_{i=1}^{V} (c_i + \alpha) \,, \tag{2}$$

where c_w is the observed count of vocabulary token with index w, V is the size of vocabulary, and α is a smoothing constant, which we set to 0.1.

To determine which tokens are unusually common or rare in *r/science*, we used a χ^2 test of independence with a Bonferroni correction. Among those that were significantly different, we identified the tokens that had the most improbably high or low observed counts by modeling them with independent binomial distributions for each token:

$$p\left(c_w^{(s)}\right) = \text{Binomial}\left(N^{(s)}, \hat{p}_w\right), \tag{3}$$

where $c_w^{(s)}$ is the observed count of token w in *r/science* and $N^{(s)}$ is the total number of tokens (uni- or bigrams) in *r/science* (posts or comments).

To summarize these findings, we report the words whose observed counts had the lowest probability according to these models, separating them into those that were used more frequently and less frequently than would be expected. We also used a similar analysis to compare the frequency of words used by transient contributors, those who only contributed once, and experienced contributors within *r/science* comments.

To further evaluate how the language of *r/science* differs from the language of other subreddits, we built a text classifier using only uni- and bigram features to classify posts and comments as either in *r/science* or not. In contrast to the language models, which show how surprising the language of other subreddits is to *r/science*, a classifier will show how easily distinguishable *r/science* language is compared to those of other subreddits. Our classifier is a support vector machine (SVM) using uni- and bigram features. We use a term frequency inverse document frequency (tf-idf) transform to account for common words throughout all posts and comments. We trained two SVMs: one for posts and one for comments. We used one versus many classification, meaning the classifiers classify posts and comments as either in *r/science* or not. We report the F_1 score for both classifiers. Because the number of non-*r/science* posts and comments vastly outweigh the number of *r/science* posts and comments, they are sampled equally to have a balanced training and test set.

RQ2: Language used by transient vs. returning authors

Past work has shown that new authors commenting or posting for the first time do not necessarily adopt the language norms of the community immediately [8] and are more likely to leave the community. This is know as an *acculturation gap* [37], i.e., the difference in language between users who only ever posted or commented once (transient users) and returning authors, and is predictive of long term engagement [8]. To analyze whether there is a defined acculturation gap for *r/science* (which would indicate new users' difficulty in adapting to the specialized language of the community), we calculated the difference in the cross entropy of posts and comments by transient contributors (i.e., those who only contribute once) vs. experienced contributors. Following Zhang et al. [37], we use experienced contributors as a proxy for community language. It is also important to note that transient users might have contributed to other subreddits or even read *r/science* posts; however, they had not posted or commented in *r/science*. We calculated the cross entropy of experienced contributors by sampling 5 comments for 50 experienced contributors, for a total of 250 comments. We then sampled 250 comments from transient contributors. The acculturation gap was the difference between these two cross entropies. We resampled experienced and transient contributors' comments for each month and 12 times on the entire year for posts. We followed this analysis with a deeper look into the common and rare words used by transient users compared to experienced users in *r/science*. This allowed us a more nuanced perspective on not only whether the language differed between transient and returning users, but also how it differed.

While the acculturation gap is useful in identifying the distinctiveness of *r/science*'s language to first time users, we sought to further investigate whether language might act as a barrier to new users by examining if those who ultimately stayed in *r/science* matched the language of the community more closely in their first contribution than those who only contributed once and then left. If so, this would suggest that language is a factor for deterring users from engaging with *r/science* beyond one post or comment. Because we only looked at data for a single year, 2018, we were unable to determine whether users contributed before this cutoff. We therefore ignored comments and posts from January and February for this analysis. Considering that over 70% of users contribute for only one month in *r/science*, it is unlikely that users who did not post in the first two months of 2018 were active before that.

We calculated the cross entropy of the 250 first posts or comments from 250 experienced authors, comparing this to 250 posts or comments from authors who only ever commented or posted once in the subreddit. We resampled for each month of comments and 12 times on the entire year for posts.

RESULTS

RQ1: The *r/science* community uses specialized language compared to other subreddits.

The cross entropy of comments and posts significantly differed across subreddits (posts: $F_{11,14388)} = 3995.661$, $p < .0001$, comments: $F_{(11,14388)} = 1615.158$), see Figure 2. *r/science* has the lowest cross entropy for both posts and comments, suggesting that there are unique language characteristics (e.g., words and phrases) in *r/science*'s posts and comments that do not occur in other subreddits. The difference holds even for those subreddits that are topically related (e.g., *r/askscience* and *r/everythingscience*).

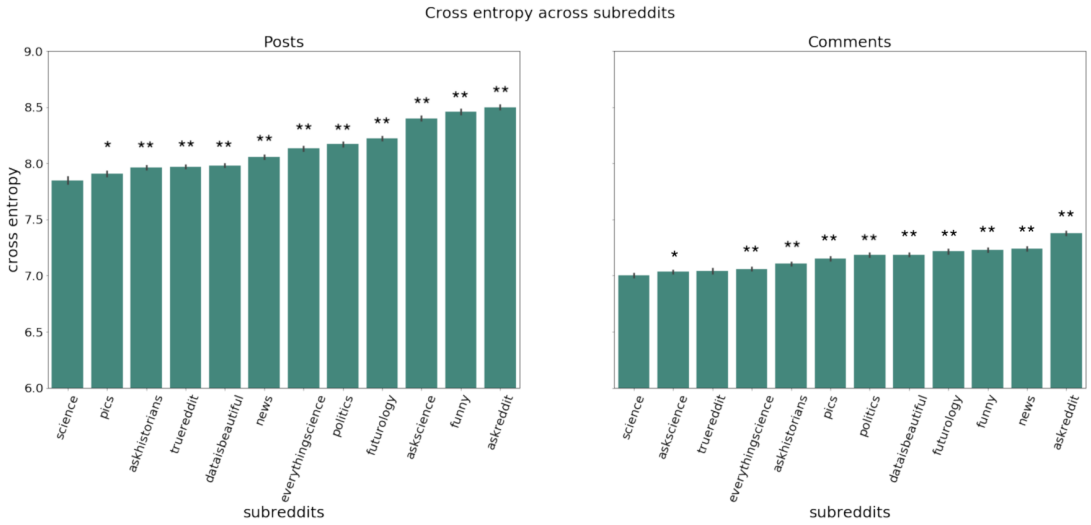

Figure 2. Cross entropies of posts and comments from experienced contributors from each subreddit, calculated using *r/science* language models. Bars show bootstrapped 95% confidence intervals. Asterisks denote t-test significance compared to *r/science*. Similar results were found using Mann-Whitney U tests for non-normal distributions. All results are corrected for multiple hypothesis testing using Bonferroni-Holm; $^{*}p < .05$, $^{**}p < .001$.

Table 2. Summary of the most unusually common and rare words in *r/science* posts and comments, calculated with χ^2 test of independence based on the frequency of the words in *r/science* relative to other subreddit posts and comments, respectively.

	Especially Common	Especially Rare
Posts	Terminology (*cells, brain, cancer, disease*)	Pronouns (*you, your, it, youve, my, i, he, me*)
	Reporting (*study, researchers, scientists, new*)	Questions (*what, whats, if, why, how, who*)
	Hedge words (*may, according, suggests, likely*)	Opinions (*serious, best, worst, like, favorite*)
Comments	Terminology (*cells, cancer, species, energy*)	Pronouns (*he, i, his, she, my, her, him, me*)
	Reporting (*study, studies, science, research*)	Politics (*trump, mueller, clinton, hillary*)
	Analysis (*factors, correlation, likely, link*)	Profanity

While it is not surprising that *r/science* has the lowest cross entropy since the language models were trained on it, it is interesting to see how other subreddits relate to *r/science* in post and comments. For example, post cross entropies are more different across subreddits than are comments. This is due to posts containing more topic words (e.g., *Trump*, *cells*) than comment text. Interestingly, *r/pics* posts have a similar cross entropy to *r/science* posts. This may be because *r/pics* also has stringent guidelines for post titles that discourage personal words in the title (e.g., no memorial posts, no posts asking for assistance, and no personal information). While *r/politics* also has stringent post title guidelines, the strong topical differences between it and *r/science* most likely contributed to its higher cross entropy.

Table 2 summarizes the words and bigrams that comprise the most improbably common and rare terms in *r/science* compared to the other subreddits for both posts and comments (ignoring conjunctions and prepositions). Looking at these words, we can see that in *r/science* posts scientific terminology, references to scientific studies, and *hedge* words (e.g., *may*, *suggests*, *likely*) are all extremely common, relative to other subreddits. By contrast, personal pronouns, question words, and expressions of opinions are extremely uncommon. Similar patterns hold for *r/science* comments, but we did not observe such an extreme use of hedge words, and the most notably underused words (besides personal pronouns) are profanity and terms related to politics (which have a high background frequency in comments in other subreddits). Looking at the bigrams reveals similar findings: science posts and comments contain more references to scientific studies (e.g., *(researchers have)*, *(study finds)*) and hedge phrases (e.g., *(according to)*, *(likely to)*)) and fewer questions and personal references (e.g., *(what is)*, *(do you)*, *(when i)*. For a list of the top 50 common and rare uni- and bigrams in *r/science*, see the appendix. While there are topical differences in some word usage comparison (e.g., *Trump* as a common word outside of *r/science*), there are also many examples of stylistic differences (e.g., hedging and impersonal language) in words and phrases. This indicates that *r/science* differs in style and topic from other subreddits, especially with hedging and impersonal speech.

The classifiers achieved mixed results for identifying posts and comments as from *r/science* or not. The comment classifier obtained a test F_1-score of 73, while the post classifier reached 84. Similar to the cross entropy findings, the classifier scores suggest that posts are easier to differentiate across subreddits than comments. Considering the minimal features used to train both classifiers (simple uni and bigram features with a tf-idf transform), the scores reflect the distinctiveness of *r/science* posts, scoring well above random chance.

RQ2: *r/science* users that don't match the community's language are more likely to leave.

The majority (57%) of users in *r/science* only post or comment once and never return. We found a pronounced and significant difference in cross entropy of these transient authors versus experienced authors in *r/science* for comments (mean = 7.37,

Table 3. Posts and comments from transient and experienced users in *r/science*. Sampled from all contributions.

	Transient Contributors	Experienced Contributors
Posts	*The definition of the kilogram might be about to change for the better!*	*Sulfur isotope has helped reveal surprising information about both the origins of life on Earth.*
	How Reddit (and the rest of the internet) is good (and bad) for you	*Negative experiences on social media carry more weight than positive interactions [...]*
Comments	*Same with adderall in my case. Whenever I'm on it im no longer constantly hungry*	*Yeah I would have wanted a control group just to confirm how fmri changes when you were just exposed to it.*
	I'm convinced that any mouse with a strong background in science could make itself immortal.	*Well, no, this would be enough to be revolutionary if you could build, say, MRI machines with it. It's much cheaper to run a fridge than to keep something chilled with liquid helium.*

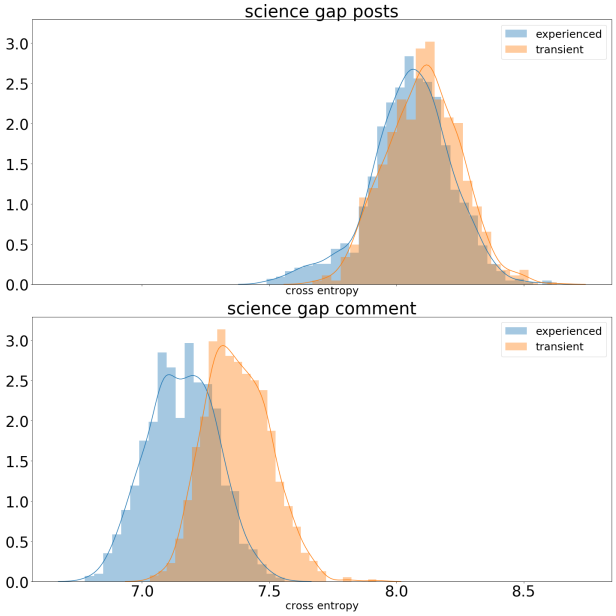

Figure 3. Sampled distribution of cross entropies between transient and experienced contributors (all posts and comments). Difference in comment cross-entropy is significant ($p < .001$) though not for posts (independent samples t-test, corrected for mutliple hypothesis testing using Bonferroni-Holm). Considering that posting is speaking to all of *r/science*, this is most likely due to people following *r/science*'s posting rules more closely when they first post than when they first comment.

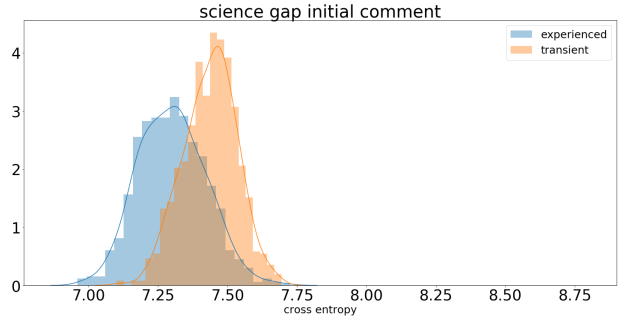

Figure 4. Sampled distribution of cross entropies between first time comments of users who leave and who stay. Difference is significant ($p < .001$).

s.d. = 0.13, vs. mean=7.16, s.d. = 0.14) ($t_{1199} = 40.85$, $p < .0001$, $d = 1.67$) and posts (mean = 8.10, s.d. = 0.15, vs. mean = 8.05, s.d. = 0.16) ($t_{1199} = 8.07$, $p < .0001$, $d = 0.33$). Figure 3 plots the distributions of the cross entropies for posts and comments.

Looking at common words and phrases in these transient user comments, we found that personal words (e.g., *i*, *my*, *feel*) are significantly more common in transient user comments than experienced user comments of *r/science*, and words discussing scientific findings (e.g., *abstract, journal, evidence*) are significantly rarer. The common and rare bigrams for transient users reflect similar differences, with personal and anecdotal phrases (e.g., *(i was)*, *(when, i)*) common while references to scientific findings (e.g., *(linked, academic)*, *(press, release)*) rare. These results mirror those found between *r/science* and other subreddits (see Table 2) suggesting that users from other popular subreddits, while possibly matching the language in these subreddits, are faced with a more pronounced language barrier in *r/science*. Table 3 provides examples of posts and comments from transient and experienced contributors on *r/science*.

r/science also stands out as having lower new user retention than the majority of other subreddits (see Figure 5), with an average of 10% (s.d. = 3.28%) of new users returning after posting or commenting for the first time in the previous month ($F_{(11,120)} = 13.818$, $p < .0001$). Interestingly, many of the subreddits related to *r/science*, such as *r/askscience* and *r/everythingscience*, have similarly low retention.

Our methods for measuring language differences did not allow quantitative comparisons of differences between transient and returning authors in *r/science* compared to other subreddits since this would have involve comparing significance values and cross entropies calculated from different language models, both of which are improper comparisons. We instead ran our word frequency tests on comments of transient users from subreddits topically related to *r/science* and with similarly low user retention: *r/askscience* and *r/everythingscience*. If the common and rare words fell into similar categories as in *r/science* (e.g., personal words and scientific findings) for transient contributors, this would suggest that the low retention in these communities is related new users failing to adapt to the same specialized language.

Common words for transient users in *r/askscience* and *r/everythingscience* included more personal words words (e.g., *i*, *my*, *your*), while scientific words (e.g., *species*, *particles*, *genetic*) were rare for transient contributors. However, there were also noticeable differences in these words compared to *r/science*: *r/askscience* contained many more question words

(e.g., *what*, *please*, *thank*), which makes sense considering the purpose of the subreddit is to ask questions. These differences fall along the differences in the purposes of the subreddits, while the similarities between these subreddits (a focus on scientific terminology and away from personal words) suggests that this type of language is more difficult for the general Reddit user to adapt to, possibly contributing to the lower user retention they share.

These differences suggest that transient users in *r/science*, those who only post or comment once, use significantly different language than those who are returning contributors in the community. To delve deeper into this difference, we explored how the language of the first post or comment of contributors in *r/science* who would return differed from the posts or comments made by transient users.

Users who end up becoming experienced contributors of *r/science* matched the language in *r/science* in their first comment more closely than those who only contributed once and then left (mean = 7.30, s.d. = .13 for experienced users vs. mean = 7.40, s.d. =.11 for transient users) ($t_{1199} = 19.03$, $p < .0001$, $d = 0.78$). Figure 4 plots this difference, showing similar distributions as found in Figure 3 for comments. We did not find this to be the case with posts; the cross entropy of experienced users' first posts was not significantly lower than the cross entropy for transient users' posts. Considering that posting is speaking to all of *r/science*, this is probably due to people focusing more on what they are writing – and following *r/science*'s posting rules more closely – when they first post compared to when they first comment. These results show that users who leave after commenting once diverge from the language of *r/science* significantly more than users who ultimately stay, suggesting that language is a factor for deterring some users from commenting in *r/science*.

Past work has shown that users who are more likely to stay in an online community write differently, such as using more collective identity words, than those who are just passing through, even in their first post or comment [13]. However, other work has also shown that users begin writing differently, such as using more collective identity words, the longer they stay in a community [27]. This highlights a debate in social computing on whether highly active users in a community are born, meaning there is something inherently different about these users, or made, meaning users are slowly drawn into a community [15]. While Figure 4 suggests that returning users in *r/science* are born rather than made when commenting, we decided to further explore this concept of born versus made by analyzing how comment cross entropy changed as a function of how long the user had been part of *r/science*. To do this, we sampled 1000 experienced users and numbered all their comments from 1 (the user's first comment in our data) to n (their last) and calculated cross entropy for all comments. We used comment number, rather than actual timestamp, as a our measure of time in the community because each user contributes to the community at a different rate, making physical time an inaccurate measure for observing change across users [8]. We ignored comments numbered greater than 50,

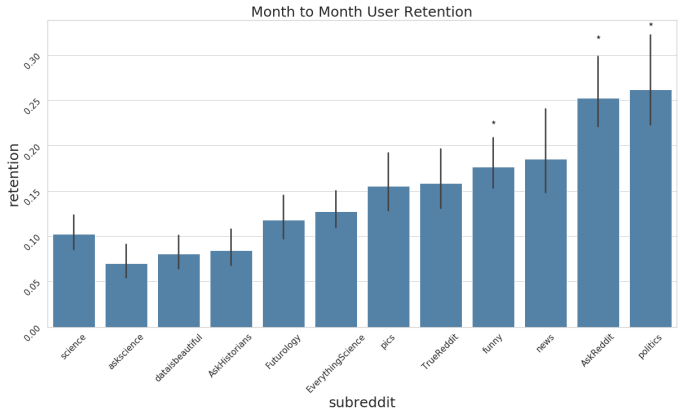

Figure 5. New user retention across all subreddits. $^{*}p < .05$ **independent samples t-test compared to *r/science*. Corrected for mutliple hypothesis testing using Bonferroni-Holm.**

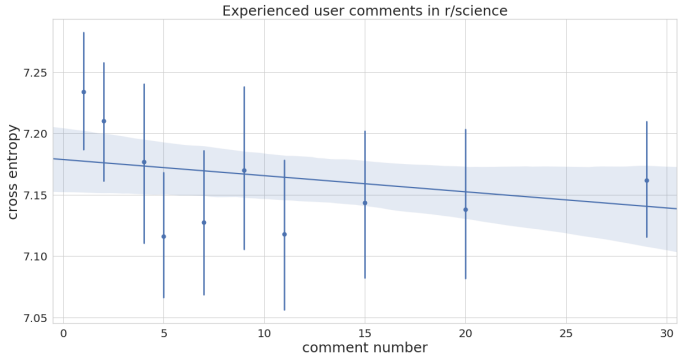

Figure 6. Cross entropy of contributors as they contribute more to *r/science*. Cross entropies are binned into 10 equal groups of comments. Correlation is significant using Spearman's rank-order correlation ($\rho = -0.027$, $p < .01$).

since this represented a tiny minority of contributors (less than half of 1%) and would encourage overfitting on a small subset.

We found that over time users will match the language of *r/science* more closely. Figure 6 plots cross entropy as a function of comment number for contributors in *r/science*. As the figure shows, there is a slight negative correlation between comment number and cross entropy ($\rho = -.027$, $p < .01$, using Spearman's rank-order correlation), showing that the longer a user contributes to *r/science*, the closer their language matches the language of other experienced contributors in the community. This is in line with prior work on socialization in online communities [8, 27] and suggests that although experienced users may begin by matching *r/science*'s language more than transient users (supporting a born hypothesis) they will also match the language of *r/science* more as they continue to comment (supporting a made hypothesis) [15].

DISCUSSION

Science communication can help stimulate awareness and improve understanding of science, creating a society that appreciates and supports science and science literacy [3]. In an ideal world, people of various walks of life would come together and talk about science. Online forums, such as the

science communication subreddit *r/science*, have the potential to make this happen. In theory, anyone with Internet access can participate on *r/science* and engage in science discussions. However, our results suggest that language use can act as a barrier for participation. We found that those who actively contribute on *r/science* frequently use scientific terminology, impersonal speech, and hedge words, such as *may*, *suggests*, or *likely*. The language used on *r/science* is distinctively different from other popular and topically related subreddits as evidenced by our analyses. Our results support past research on specialized language as a hurdle to science communication in general [29] and extend it by showing how specialized language gatekeeps in social discussions of science. This works provides an exciting step towards exploring the gatekeeping effect of language in scientific communication, both on *r/science* and in other media such as blogs and news articles, making it more equitable and accessible.

While the specialized language on *r/science* may be a natural extension of the scientific expertise community members have [17], our findings suggest that not everyone is able or willing to learn *r/science*'s specialized language. Those who contribute with strongly diverging language from *r/science* might still passively consume the content or contribute through upvotes and downvotes of content, but ultimately refrain from posting or commenting again. The result is especially noteworthy because it highlights a lost potential in encouraging those who posted or commented once to become active community members. If people's language used in a post and comment is indicative of diversity (which prior work suggests [26]), then this also signifies a lost opportunity to involve a broad range of people with varying scientific familiarity.

Interestingly, *r/science*'s community rules on the front page (Figure 1) reflect much of this specialized language in explicit language norms. For example, *r/science*'s rules against sensationalized titles and personal anecdotes[4] have clear parallels to the abundance of hedge words and rarity of personal and subjective words. In addition, existing posts and comments act as a descriptive influence of norms [7] by providing language examples that have been accepted by the community.

One interpretation of our results is that the community rules are effective and functioning as intended. First, *r/science*'s specialized language is clearly geared toward maintaining rigorous scientific discussion in the community [17]. Second, it could be that the language of *r/science* promotes stability in readership and trust among passive users; by having a high bar for who gets to talk, people feel most of what they hear is worth listening to. This can encourage a readership that lurks but does not contribute, trusting in those who speak the language to contribute to the discussion. We see some evidence of this in the massive number of subscribers in *r/science* (over 21 million, fifth overall on Reddit at the time of writing) versus an order of magnitude fewer posts and comments compared to other subreddits with similar subscriber counts.

However, another perspective is that there are opportunities to help broaden participation and socialize newcomers. While it may not be necessary for everyone to have the same knowledge about or interest in science, science communication should be accessible to everyone [3]. Our findings indicate that simply posting community rules on the front page has not been sufficient in guiding first time and transient contributors to use the specialized language expected on *r/science*. The *r/science* community can use these findings and our methodology to explore ways of socializing newcomers without sacrificing community norms, such as by identifying and defining specialized terms in the guidelines or moderator posts.

We noticed qualitative evidence of this same tension between maintaining rigorous standards and broader participation in *r/science* comments during our period of analysis (January to December 2018). When searching for comments that contained 'this subreddit' in *r/science* comments, we found that there was a small number of comments on both sides of this debate. While some comments bemoaned the increase in clickbait titles and pop media, others criticized the practice of deleting less serious contributions and those from a broader audience. We look forward to presenting our results to *r/science* moderators as a way of furthering this discussion and identifying possible ways to support new users without sacrificing the standards of *r/science*. Our findings and methods provide the exciting first step for tools to support users in their first contributions to scientific communication communities and ways to lower the barrier of participation in the first place, making scientific communication more equitable and accessible. Below we discuss opportunities to make this happen.

Design Implications

Acculturating new users to the language norms of r/science.
One way of encouraging broader participation while maintaining *r/science*'s specialized language is a more effective way of welcoming newcomers to the community. Posting guidelines can be helpful in enforcing norms, but if norms are difficult for newcomers to understand or follow, they will leave. Providing feedback on first contributions can help new users follow community norms more effectively [10]. This feedback has previously been given by veteran users willing to take time and help new contributors, but our findings suggest that automatic measures can capture and convey language norms in a community. Our approach in this paper of using language models and word frequency distributions provides an effective way of identifying diverging language and highlighting common and rare words for the community. For example, we identified common hedging words that users who are more likely to stay in *r/science* will use. Language models like ours can provide automatic, just in time feedback for posts and comments and offer a scalable way of helping new users contribute.

Reduce or explain scientific jargon
While jargon is important in communicating to other researchers focused on the same problem, it can overwhelm users with less domain experience. Scientific blogs for readers outside of the research community often take great care to reduce or explain scientific jargon [30], which can keep them faithful to the research without overwhelming their audience. Our methods of identifying common and rare words can help *r/science* moderators and scientific writers identify jargon

[4] https://www.reddit.com/r/science/wiki/rules#wiki_submission_rules

that can act as a linguistic barrier to new users, constructing a guide to new users on commonly used scientific words and their plain English explanations. Taking inspiration from tools that encourage simpler language (e.g., xkcd's Simple Writer [25]), language models like ours can identify unusual words to reduce jargon when users are posting or commenting, lowering the linguistic barrier of entry while maintaining accuracy in *r/science*. Going further, language-aware writing technologies can support writers more than just highlighting words, potentially surfacing automated questions or edits for improving clarity or audience reception.

Scientific discourse with less strict language norms

One other possibility to encourage people with diverging language from actively contributing over a longer period of time could be to provide opportunities for reading and contributing posts and comments that are less strict in their language norms. *r/science* could offer an extra space for such discourse where moderators could follow up on posts that compromise rigor. While this would essentially create a subcommunity, an advantage of this approach is that people might not feel alienated by incomprehensible content. *r/science* does have a sister subreddit, *r/everythingscience*, which is ostensibly more open to broader discussion; however, the posted guidelines are similar to *r/science*'s, with one main difference being the post does not need to reference a peer-reviewed paper from the last 6 months. This suggests that the language barriers between the two might not be so different. *r/everythingscience* could experiment with relaxing some language norms to encourage broader participation.

Rather than offering a separate subcommunity, *r/science* could also offer different interface views. One view could show all posts and comments together but color code those that were intended for a broader audience. Users could choose whether they prefer to see only posts and comments that strictly follow *r/science* guidelines versus a mixed view.

LIMITATIONS & FUTURE WORK

We focused only on first time authors as newcomers; however, work has shown that users often lurk in a community, observing acceptable behaviour and norms, before contributing [10]. Lurkers might only contribute occasionally (e.g., once in a year), including them in our classification of transient contributors while in actuality they may be fully acculturated to the community. The fact that we found differences in language between transient and returning members suggests that lurkers who post occasionally form a minority of transient users, or this initial post is a catalyst for more active contribution. We also aggregated across subreddits to study how language differed at the community level, treating each subreddit separately. An exciting extension of this work is to explore how the same contributor's language differs depending on what community they are in.

Most automatic measures of language style, including the language models we used, count relative frequency of words, making it difficult to disentangle style from topic [35, 37] or what people are talking about versus how they are talking about it. An exciting future direction of this work is to develop measures of language style robust to topical variation.

CONCLUSION

This paper contributes a comparison of the language used in *r/science*, a subforum on Reddit that is focused on science communication, to 11 other large or topically related subreddits. We showed that *r/science* uses a specialized language distinct from the other subreddits and that transient contributors to *r/science* (those that contribute only once) do not match this specialized language compared to contributors who ultimately stay, even when comparing their first comments. These results indicate that *r/science*'s specialized language acts as a gatekeeper to the community, discouraging a broader audience from contributing. Our findings add to the ongoing discussion of how to involve a diverse public in scientific discourse by providing methods for reducing linguistic barriers of entry to one of the largest forums of public scientific discussion.

DATASET AND TOOLS

We make available our datasets and Python code used for analysis at https://github.com/talaugust/rscience_language.

ACKNOWLEDGMENTS

This work was partially funded by NSF award 1651487. We thank the ARK lab and Justine Zhang for providing valuable feedback and support. We also thank the reviewers for their time and input.

REFERENCES

[1] Dominique Brossard and Dietram A. Scheufele. 2013. Science, new media, and the public. *Science* 339, 6115 (2013), 40–41.

[2] Moira Burke and Robert Kraut. 2008. Mind your Ps and Qs: the impact of politeness and rudeness in online communities. In *Proceedings of the 2008 ACM Conference on Computer supported Cooperative Work*. ACM, 281–284.

[3] Terry W. Burns, D. John O'Connor, and Susan M. Stocklmayer. 2003. Science communication: a contemporary definition. *Public Understanding of Science* 12, 2 (2003), 183–202.

[4] Justine Cassell and Dona Tversky. 2005. The language of online intercultural community formation. *Journal of Computer-Mediated Communication* 10, 2 (2005).

[5] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. In *Proceedings of the ACM on Human-Computer Interaction*.

[6] Stanley F. Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language* 13, 4 (1999), 359–394.

[7] Robert B. Cialdini, Raymond R. Reno, and Carl A. Kallgren. 1990. A focus theory of normative conduct: recycling the concept of norms to reduce littering in

public places. *Journal of Personality and Social Psychology* 58, 6 (1990).

[8] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 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.

[9] Casey Fiesler, Jialun "Aaron" Jiang, Joshua McCann, Kyle Frye, and Jed R. Brubaker. 2018. Reddit rules! Characterizing an ecosystem of governance. In *Proceedings of International AAAI Conference on Web and Social Media (ICWSM)*.

[10] Denae Ford, Kristina Lustig, Jeremy Banks, and Chris Parnin. 2018. We don't do that here: How collaborative editing with mentors improves engagement in social Q&A communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

[11] Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. 2013. The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist* 57, 5 (2013), 664–688.

[12] Aaron Halfaker, Aniket Kittur, Robert Kraut, and John Riedl. 2009. A jury of your peers: quality, experience and ownership in Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*. ACM.

[13] William L Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Loyalty in online communities. In *Eleventh International AAAI Conference on Web and Social Media*.

[14] Per Hetland. 2014. Models in science communication: formatting public engagement and expertise. *Nordic Journal of Science and Technology Studies* 2, 2 (2014), 5–17.

[15] Shih-Wen Huang, Minhyang Mia Suh, Benjamin Mako Hill, and Gary Hsieh. 2015. How activists are both born and made: An analysis of users on change. org. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 211–220.

[16] Aaron Jaech, Victoria Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. 2015. Talking to the crowd: What do people react to in online discussions?. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2026–2031.

[17] Ridley Jones, Lucas Colusso, Katharina Reinecke, and Gary Hsieh. 2019. r/science: Challenges and opportunities for online science communication. (2019).

[18] Robert E. Kraut and Paul Resnick. 2012. *Building Successful Online Communities: Evidence-based Social Design*. MIT Press.

[19] William Labov. 1973. The linguistic consequences of being a lame. *Language in Society* 2, 1 (1973), 81–115.

[20] William Labov. 2006. *The Social Stratification of English in New York City*. Cambridge University Press.

[21] Trevor Martin. 2017. community2vec: Vector representations of online communities encode semantic relationships. In *Proceedings of the Second Workshop on NLP and Computational Social Science*. 27–31.

[22] J. Nathan Matias and Merry Mou. 2018. CivilServant: Community-led experiments in platform governance. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM.

[23] James Milroy and Lesley Milroy. 1978. Belfast: Change and variation in an urban vernacular. *Sociolinguistic Patterns in British English* 19 (1978), 19–36.

[24] Jonathan T. Morgan and Anna Filippova. 2018. 'Welcome' changes?: Descriptive and injunctive norms in a Wikipedia sub-community. *Proceedings of the ACM on Human-Computer Interaction* 52 (2018).

[25] Randall Munroe. 2015. A Thing Explainer word checker. (2015). `https://blog.xkcd.com/2015/09/22/a-thing-explainer-word-checker/`

[26] Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics* 42, 3 (2016), 537–593.

[27] Dong Nguyen and Carolyn P. Rosé. 2011. Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Languages in Social Media*. 76–85.

[28] Nigini Oliveira, Michael Muller, Nazareno Andrade, and Katharina Reinecke. 2018. The exchange in StackExchange: Divergences between Stack Overflow and its culturally diverse participants. *Proceedings of the ACM on Human-Computer Interaction* (2018).

[29] Pontus Plavén-Sigray, Granville James Matheson, Björn Christian Schiffler, and William Hedley Thompson. 2017. The readability of scientific texts is decreasing over time. *eLife* 6 (2017).

[30] Mathieu Ranger and Karen Bultitude. 2016. 'The kind of mildly curious sort of science interested person like me': Science bloggers' practices relating to audience recruitment. *Public Understanding of Science* 25, 3 (2016), 361–378.

[31] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* (2019).

[32] Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 641.

[33] Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Proceedings of the Seventh International Conference on Spoken Language Processing*.

[34] Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

[35] Trang Tran and Mari Ostendorf. 2016. Characterizing the language of online communities and its relation to community reception. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 1030–1035.

[36] Debbie Treise and Michael F. Weigold. 2002. Advancing science communication: A survey of science communicators. *Science Communication* 23, 3 (2002), 310–322.

[37] Justine Zhang, William L Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.