No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities

Cristian Danescu-Niculescu-Mizil
Stanford University
Max Planck Institute SWS
cristiand@cs.stanford.edu

Robert West
Stanford University
west@cs.stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

Christopher Potts
Stanford University
cgpotts@stanford.edu

ABSTRACT

Vibrant online communities are in constant flux. As members join and depart, the interactional norms evolve, stimulating further changes to the membership and its social dynamics. Linguistic change—in the sense of innovation that becomes accepted as the norm—is essential to this dynamic process: it both facilitates individual expression and fosters the emergence of a collective identity.

We propose a framework for tracking linguistic change as it happens and for understanding how specific users react to these evolving norms. By applying this framework to two large online communities we show that users follow a determined two-stage lifecycle with respect to their susceptibility to linguistic change: a linguistically innovative learning phase in which users adopt the language of the community followed by a conservative phase in which users stop changing and the evolving community norms pass the user by.

Building on this observation, we show how this framework can be used to detect, early in a user's career, how long she will stay active in the community. Thus, this work has practical significance for those who design and maintain online communities. It also yields new theoretical insights into the evolution of linguistic norms and the complex interplay between community-level and individual-level linguistic change.

Categories and Subject Descriptors: J.4: [Computer Applications]: Social and behavioral sciences

Keywords: linguistic change; community norms; conventions; user abandonment; lifecycle; reviews; social influence; language

1. INTRODUCTION

"It takes a long time to become young."

—Pablo Picasso

Online communities, such as online discussion forums or product review websites, are constantly evolving. Norms of interaction change over time, from domain-specific jargon [14] to conventions for content attribution [20]. When new members join, they can adapt to existing community norms, but can also push them in new directions. Long-time members may adapt to these new norms or they may be innovators themselves, setting new trends. Other users may not react to changes, sticking to their previous styles.

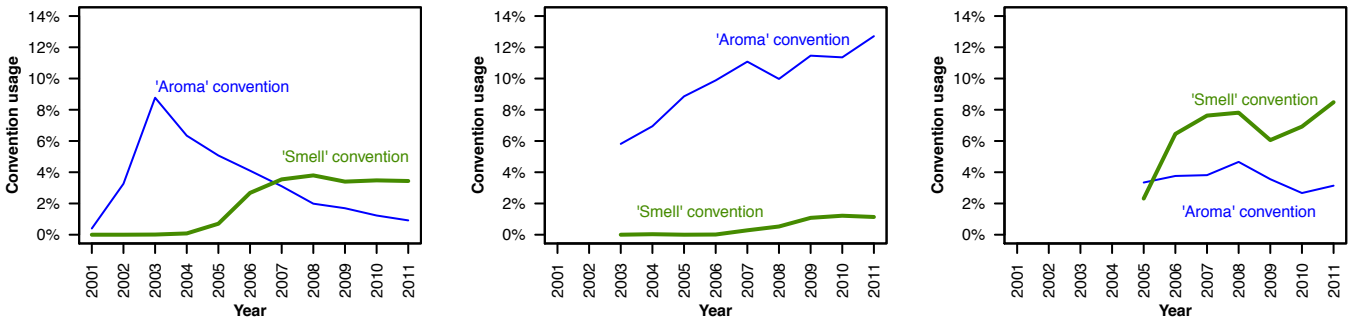
WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2035-1/13/05.

Understanding the complex ways in which the interactional styles of members and their communities jointly evolve over time is essential for building and maintaining vibrant online communities. Such conventions affect users' decisions to actively participate in the community and to contribute content [1, 31, 36] and are closely related to the dynamics of member arrival and departure.

In this paper, we study these issues through the lens of the language used by individual community members and by the community at large. In particular, we investigate *language change* in the sense of [11, 22, 25, 40, 43]: linguistic innovation originating in a sub-group that becomes accepted as the norm through a process of conforming. Such innovations facilitate individual expression and help to create tight-knit sub-cultures. At the same time, the process of conforming fosters cohesiveness within the group as a whole. The evolving norms are thus a window into the broader process of co-evolution of members and communities, serving to differentiate newcomers from long-time members and conveying information about the degree to which members remain engaged in the community.

Summary of main contributions. We propose a framework for tracking linguistic change and for understanding how specific users react to evolving community norms at different stages of their lives within their communities. We use this framework to study two large, active online communities: RateBeer and BeerAdvocate. Both sites are built around members evaluating and discussing beer. They are excellent settings in which to study linguistic change: both are more than a decade old, which gives us a long time window to work with; and both have extremely active memberships (it is common for individuals to have written more than 100 reviews). Moreover, both communities have also developed a rich set of conventions and terminology.

In applying our framework to this data, we show that users follow a determined lifecycle with respect to their susceptibility to linguistic change: early in her career, a user becomes increasingly receptive to the norms of the community up to about one third of her eventual lifespan, when she reaches a maximum synchrony with the language of the community (we will call this early period *linguistic adolescence*); from that point on, a gap between the user's language and that of the community forms and increases until the moment she abandons the site. We show that this increasing gap is explained by the user ceasing to respond to changes in community norms: because the language of the community is constantly evolving a user that is unreceptive to this change will appear as fathering from the community.



(a) 'Aroma' was the dominant convention by 2003, but it was supplanted by 'S' (for 'Smell') around 2007.

(b) Users who joined in 2003 hung on to the 'Aroma' convention of their youth.

(c) Users who joined in 2005 were more receptive to the emerging 'S' norm.

Figure 1: Example of community and user evolution in BeerAdvocate: one norm for referring to the smell of a beer gave way to another, with different effects on different users depending on when they joined the community.

This pattern is surprisingly consistent, with users following the same lifecycle pattern regardless of how much time or effort they actually spent in the community. Long time contributors and short term users follow the same pattern, with the latter evolving at an accelerated pace. Crucially, this means that at the moment they depart from the community most users are linguistically conservative.

Building on these observations, we show that a user's patterns of linguistic change can be harnessed to make predictions about her membership in the community (Section 4). Inspired by our empirical analysis, we design features that capture the speed at which a member conforms to community norms along with other linguistic change patterns, and use them in simple machine learning models. We find that, using solely the language of the first few initial reviews, we can predict a user's total lifetime in the community. Our models give significant performance improvement over a baseline model that is based on traditional activity based features [9]. Our results thus have practical significance for designers and maintainers of online communities, who can use them to detect early in a user's career how long she will stay active in the community.

Implications for social networks and sociolinguistics research. This work also yields new theoretical insights into the evolution of linguistic norms and the complex interplay between community-level and individual-level linguistic change, addressing important open questions in the social network and sociolinguistic literature. There is extensive research covering the evolution of online communities [21, 28], the evolution of community members [2, 10, 16, 29, 32], and their participation patterns [1, 3, 26, 30, 31, 48] (see Section 6 for a discussion). But little is known about the interplay between user-level evolution and the evolution of the community at large, an issue which is the crux of this work.

This interplay is also a central open research question in linguistics. Much sociolinguistic research has relied on the *adult language stability assumption*: under the critical assumption that individuals' speech patterns are largely fixed by early adulthood, older speakers' language can be employed as a proxy for the linguistic state of the community at an earlier stage [11, 12, 22, 23, 37, 42]. However, studies have also shown that this assumption can fail to hold. For instance, individuals might change their language as they age, while the community as a whole remains stable; or a change might be simultaneously adopted by all members of the community, or just by older members [25, 39, 40, 46, 47]. Distinguishing among scenarios like these has proved to be a fundamental challenge in sociolinguistics research [43]. Our framework opens up new avenues for the studying of these issues in highly dynamic communities, and our results show how the adult language stability assumption and other theoretical models of linguistic change apply to online setings (see Section 5 for a detailed discussion).

Linguistic change: An example. The bulk of this paper is dedicated to developing a framework for tracking the aggregate effects of numerous, ever-evolving linguistic changes. Not all of these changes are intuitively accessible; like the phonological effects primarily studied by linguists working in offline communities, the changes are often difficult to identify and characterize individually. Thus, in what follows, we rely on quantitative evidence and high-level evaluations to assess our framework. Nonetheless, many of the linguistic changes at work in our data are highly intuitive and easy to discern. It is illustrative to review one of them, before we move to studying them at a more abstract level. The themes of this example play out many times over in experiments to come.

Because our communities are built around beer, the discussion frequently turns to assessing various aspects of the beer-tasting experience, so this is a locus of linguistic change at the lexical level. One prominent example concerns smell. Over the life of the BeerAdvocate community, there were two prominent conventions used to introduce discussions of smell: *Aroma* and *S* (short for 'Smell'). Figure 1(a) summarizes the basic trend for this linguistic variable: the *Aroma* convention (blue, thinner line) rose quickly in popularity between 2001 and 2003, when it reached its peak. Around 2003, the *S* convention (green) began its rise in popularity. The *Aroma* convention lost ground quickly and was soon overtaken by *S*.

Figures 1(b) and 1(c) show that this linguistic change affected old users differently than it affected new ones. Users who joined the site in 2003, at the height of the *Aroma* boom, were very unlikely to switch to *S* (Figure 1(b)). Indeed, they make hardly any use of *S* (green), and even increase their use of *Aroma* (blue), possibly as a reaction to the encroaching norm and the social changes it potentially signals [47]. The picture is very different for users who joined in 2005, when *S* was taking off. Figure 1(c) suggests that these new users are drivers of this change; their *S* usage rises sharply while their *Aroma* usage starts and remains low.

We turn now to defining our framework for tracking such changes systematically, seeking to use it to understand the social dynamics of a community and connect them with individual members' behaviors.

2. EXPERIMENTAL SETUP

In the following we first describe our dataset and then proceed to discuss some of the details of the methodology used for our analyses. In particular, here we focus on specific issues pertaining to using language models to model linguistic change in longitudinal data, leaving the bulk of the framework description to Section 3.

Community data. The framework presented in this paper is targeted at large and active online communities, where individuals interact through written text visible to all members of the community. For the purpose of this study we will employ data from two large beer review communities (BeerAdvocate and RateBeer).[1] In both communities users provide ratings accompanied by short textual reviews of more than 60,000 different types of beer. As we argue next, BeerAdvocate and RateBeer exhibit multiple features that make them suitable for the analysis of linguistic change. Statistics of the two datasets are given in Table 1.

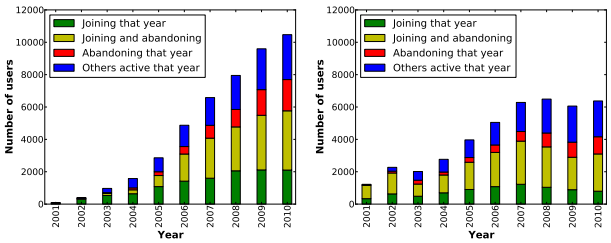
	BeerAdvocate	RateBeer
Number of posts	1,586,614	2,924,127
Number of users	33,387	29,265
Users with more than 50 posts	4,787	4,798
Median number of words per post	126	54
Median number of sentences per post	9	5

Table 1: Statistics of BeerAdvocate and RateBeer.

We crawled a complete set of reviews for BeerAdvocate and RateBeer all the way back to the inception of the site [33], spanning a period of more than 10 years—from 2001 until 2011. Thus, one of the main advantages of our datasets is the availability of the entire community history; this not only means that they provide complete longitudinal information for each user, but also that linguistic conventions can be tracked back to their initial introduction (we have already discussed two examples of such conventions in the Introduction).

Another reason why these datasets are suitable for our purposes is that users commonly contribute substantially to the community (e.g., more than 4,700 users wrote at least 50 posts) and this becomes particularly important when tracking the patterns of linguistic change over the lifespan of a user. Furthermore, since both communities were active for over a decade, we can observe multiple generations of users simultaneously, and therefore discard external effects. Lastly, these communities are united by a very specific purpose—the appreciation of beer—which makes for a fertile environment for linguistic innovation.

User lifespan. Over the course of 10 years the BeerAdvocate and RateBeer communities have evolved both in terms of their user base as well as ways in which users review and discuss beer. This presents us with an unprecedented opportunity to study linguistic change over users' entire lifespans, from the moment they joined the community—which we define as the time of their first post[2]—to the moment they abandon the community. We consider that a user abandoned the community if she did not contribute any post for at least one year. In all experiments involving a user's complete lifespan we ignore users that have posted on or after January 2011 in order to enforce this policy (the last month covered by our datasets is December 2011). Figure 2 shows a breakdown of active users showing the number of users joining and abandoning the two communities each year, revealing their highly dynamic user bases.



(a) BeerAdvocate (b) RateBeer

Figure 2: Change in user base: Breakdown of active users each year. From bottom up: users that joined the community that year (and did not abandon the same year), users that joined and abandoned the community that year, users that abandoned the community that year (and did not join the same year) and other active users. (a) BeerAdvocate; (b) RateBeer.

Snapshot language models. In order to relate a user's language to that of the community, we require a reliable model of the linguistic state of the community at various points in time. For that purpose we use of a series of *snapshot language models,* one for each month in the life of the community. These are bigram language models with Katz back-off smoothing[3] [17] estimated from a held-out subset of posts from each month (posts which are not used in any of the subsequent analyses).[4] Given a post p, we can then quantify how surprising its language is with respect to the language the community was using at the point in time when it was written by calculating p's cross-entropy according to the snapshot language model $\text{SLM}_{m(p)}$ of the month $m(p)$ in which the post was uttered; we write this as

$$H(p, \text{SLM}_{m(p)}) = -\frac{1}{N}\sum_i \log P_{\text{SLM}_{m(p)}}(b_i),$$

where $b_1, ..., b_N$ are the bigrams making up p and $P_{\text{SLM}_{m(p)}}(b_i)$ is the probability of the bigram b_i under the snapshot language model of the post's month $m(p)$. Higher cross-entropy values indicate posts that deviate the most from the linguistic state of the community at that particular point in time.

Controlling for length effects. Longer posts inherently have larger cross-entropy and are more likely to contain elements of linguistic innovation. Currently there is no consensus on a reliable method for normalizing entropy measures in order to completely account for length effects. In order to ensure that our results are not affected by such effects, we only consider for our analysis the first $k = 30$ words of each post (unless otherwise mentioned). We

[1]The data is publicly available at http://snap.stanford.edu/data/

[2]We do not consider passive members to be linguistically active. Studying the eventual effects of lurking on language use is an interesting direction for future work.

[3]We smooth the unigram back-off distribution using Laplace (additive) smoothing with a smoothing parameter of 0.2.

[4]The held-out sample of posts comes from random selection of 500 users active that month, each user contributing exactly 2 posts to the sample. This way all language models are trained on the same amount of data each month (i.e., 1000 posts), and no user is overrepresented. None of the sampled posts used to train the snapshot language models are ever used in any analysis (or prediction task). All analysis involving the use of snapshot language models will be restricted to the years 2004 through 2011, such that sufficient data is available for training.

experimented with various values of k and found our results to be stable across multiple choices of k.

3. USER LIFECYCLE AS REVEALED BY LINGUISTIC CHANGE

We proceed by describing a framework for tracking linguistic change in online communities and by discussing the insights this framework offers on the interplay between user and community evolution. In this section, we investigate some of the basic principles that govern the processes at work. We group this analysis into two parts: (1) user-level evolution and community-level evolution (Section 3.1); and (2) the interplay between these two as revealed by the way users react to community-level linguistic changes at different stages of their community life (Section 3.2). In these two parts, we identify some basic and recurring phenomena that we will show to be useful when developing techniques for a prediction task with high practical importance—detecting, early in a user's career, how long she will stay active in the community (Section 4). We will then discuss how the observed behavioral patterns relate to, and bring new insights into, issues central to the sociolinguistic literature on linguistic change (Section 5).

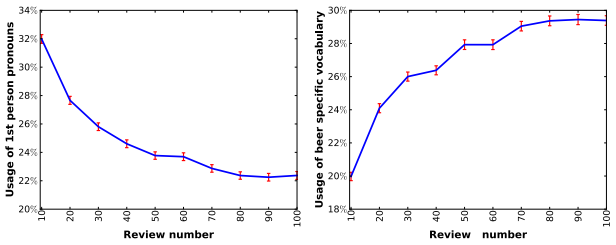
3.1 User-level and community level change

Online communities are dynamic entities, with a constantly changing user base. As such, these entities can not simply be defined as the sum of their active users. For example, out of the 662 users that joined the BeerAdvocate community in 2003, eight years later in 2011 only 16% were still active; however, the other 84% that were gone by 2011 helped define the collective identity of the community as it stands today, with linguistic norms outliving their forefathers. From the perspective of linguistic change, it is non-trivial to separate changes that occur at community-level and those that happen at the individual user-level. We start our investigations by discussing these two levels separately.

User-level change. After joining a community, a user will likely change her interaction and communication patterns. Rather than capturing specific individual changes in user behavior we are interested in capturing the overall change in user's language over time. However, for the sake of concreteness we provide examples of two specific examples of individual user change. Figure 3(a) shows how user's usage of singular first-person pronouns (*I*, *me*, *mine*, *myself*) decreases as the user contributes more reviews. This phenomenon might be attributable to a user's increasing identification with the community [5, 41]. In the same spirit, Figure 3(b) shows an increase in usage of beer specific vocabulary as the members gain experience in the community. As the user spends more time with the community, they adopt and start using the specific language of the community.

Community-level change. In a way similar to the evolution of individual users whole communities also evolve. Despite the constant change of community's user base, the community also forms an entity with a linguistic trajectory of its own. We have already discussed the example of the two community-level changes in the Introduction: the *Smell* and *Aroma* conventions that rise and fall over time (Figure 1(a)).

Figure 4 shows another example of community-level language change: The BeerAdvocate community becomes more "fruity" (using a vocabulary of 100 fruit words) over time. Even when we control for a change in the distribution of the products reviewed by macro-averaging by product, and by considering only those which



(a) First person sing. pronouns (b) Beer specific vocabulary

Figure 3: Examples of user-level language change: (a) Percentage of posts containing first person singular pronouns; (b) Percentage of reviews using specialized beer vocabulary (*retention*, *carbonation*, *lacing*, etc.). The first 100 posts of all users that contributed at least 100 posts over their lifespan are considered (so each user is represented exactly once in every bin). Results for BeerAdvocate are shown here; same trends hold for RateBeer. Throughout this paper, error bars indicate standard error estimated by bootstrap resampling [18].

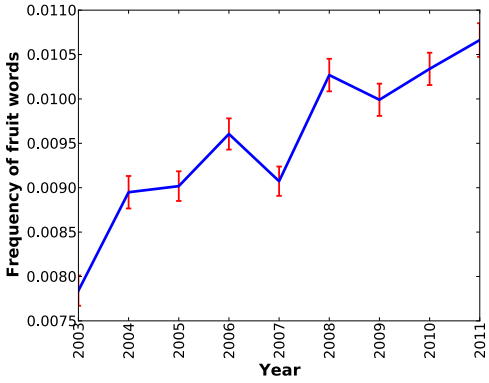
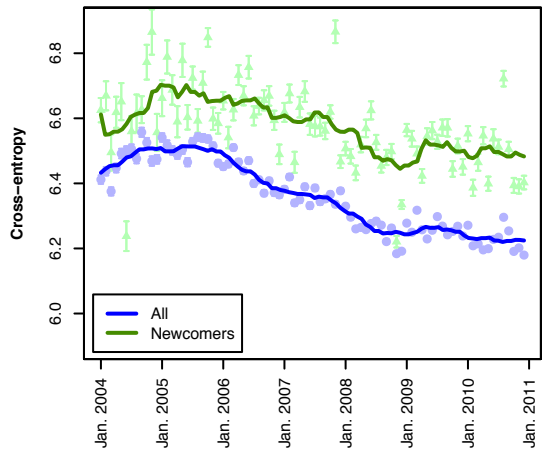


Figure 4: Example of community-level change: The usage of fruit words (e.g., *peach, pineapple, berry*) increases on BeerAdvocate. (Same trend holds for RateBeer.)

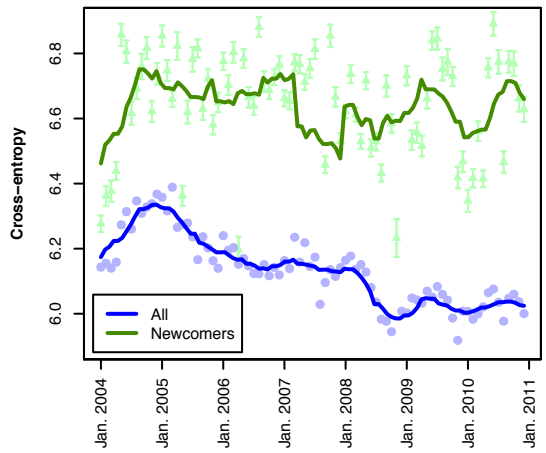
were reviewed each year, we observe a significant increase in the usage of fruit words to describe the taste and feel of a particular beer. This means that while the products are not changing the community's language to describe them changes over time.

Next we also examine the community change at a more aggregate level. In Figure 5 (blue, round markers) we examine the cross-entropy of each month's posts, according to the snapshot language model of the respective month. We make several observations. First, the cross-entropy fluctuates over time, which means that the community language is not static but it also evolves on an aggregate level. Second, the cross-entropy decreases as the community ages, which means that the posts written in the later years employ more predictable language. Moreover, this decrease in cross-entropy also suggests that evolving community norms can lead to group cohesiveness and the creation of a collective identity. Last, we also note that users that just joined use language that is less predictable and thus less aligned with the current state of the community (green, diamond markers).

The question thus arises: what is the trajectory of a user's adaptation to the community norms as she transitions from being a newcomer to being an established member of the community. We investigate this next.



(a) BeerAdvocate



(b) RateBeer

Figure 5: Example of community-level change: Predictability of each month's language, calculated as the average cross-entropy of that month's posts according to snapshot language model of the respective month (blue, round markers; error bars are very tight, sometimes not visible). Lower values mean 'easier to predict'. Compare with the predictability of the language used by users joining that month (green, diamond markers). (a) BeerAdvocate; (b) RateBeer.

3.2 User lifecycle

We now transition from exploring user-level and community-level linguistic change in isolation to the main goal of the present work: Building a framework for understanding the interaction between these levels of change. In particular, we focus on analyzing a user's susceptibility to react to the evolving norms of the community at different stages of her community life.

User life-stage. One of the challenges for an analysis that operates at both the community and the individual user level is the relativity of time. Unlike the offline settings where traditional studies of linguistic change were conducted, in online setting individuals interact with the community at very different rates. Moreover, online users have vastly different lifespans, ranging from one day to an entire decade. Therefore, it is important to identify a user's stage in their community-life in a way that allows comparison across users with different activity levels and lifespans.

To this end we define the *life-stage* of a user as the percentage of posts the user has already written, out of the total number of posts the user will ultimately write before abandoning the community.[5] Thus, a life-stage of 0% corresponds to *birth*—the moment the user joined the community—and a life-stage of 100% corresponds to *death*—the moment the user leaves the community.

User's distance from the language of the community. Another key element of the proposed framework is the ability to measure a user's reaction to linguistic change at a given stage in her community-life. In the following we use several measures for linguistic change, each of them providing different perspectives on the phenomenon. We start by quantifying the extent to which a user is in tune with the community's norms by employing the snapshot language models defined in Section 2.

In Figure 6 we plot the average cross-entropy of a user's posts at different life-stages according to the snapshot language model of the months in which the respective posts were written. Observing the evolution of cross-entropy over the users' lifespan we notice that, in both communities under study, users follow a determined lifecycle: When users join, their language is far from that of the community[6] (high cross-entropy) and then users gradually approach the current language of the community (decreasing cross-entropy); interestingly, after about a third of users (ultimate) lifespan, their language starts to again distance itself from that of the community. It appears as if a user's language falls out of tune with that of the community before she abandons the community.

Since communities as well as individuals simultaneously evolve, it is not clear whether the change in cross-entropy we just described is the result of the user actively changing her language towards (and then away) from that of the community or, on the contrary, the result of the evolving community norms getting closer (respectively away) from a static user. Thus, the increase in cross-entropy in the end stage of user's lifetime could be explained by two competing hypotheses:

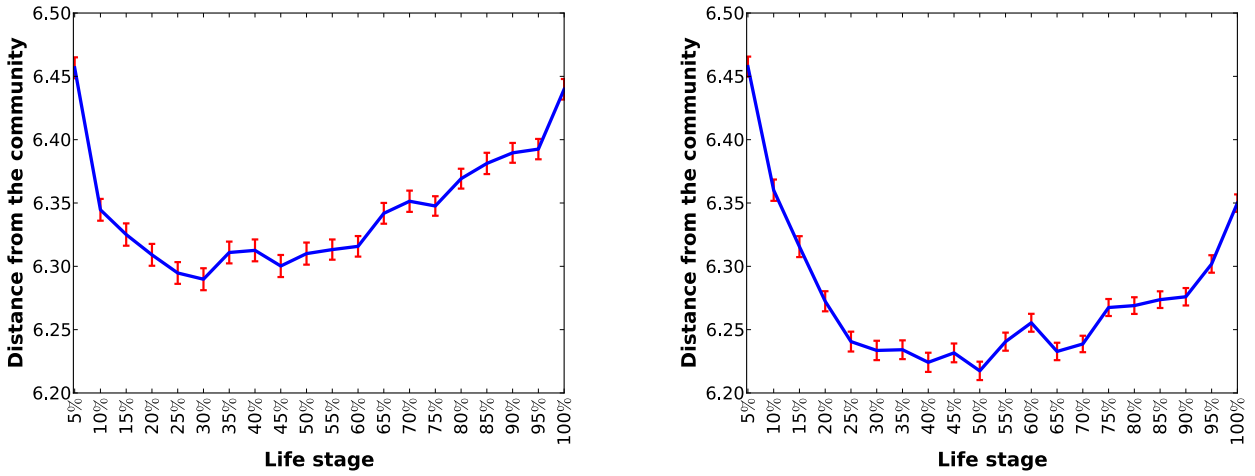
- The user is moving away from the community by starting to use language that is foreign to the current state of the community.
- The user stops adapting her language to the community and gets out of tune with the changing community.

In order to tease these two hypotheses apart, we measure how much a user's language changes with respect to her own past language at each stage of her life. More precisely, we compare the lexical overlap between each post and the previous 10 posts written by the same user according to the Jaccard similarity coefficient[7]. Figure 7(a) shows that on average users increasingly stabilize their language for the first third of their lifespan (henceforth *linguistic*

[5] In order to keep a meaningful interpretation to this fractional measure, in all experiments that involve it we ignore users with less than 50 posts. However, the same qualitative results hold if this limit is not enforced.

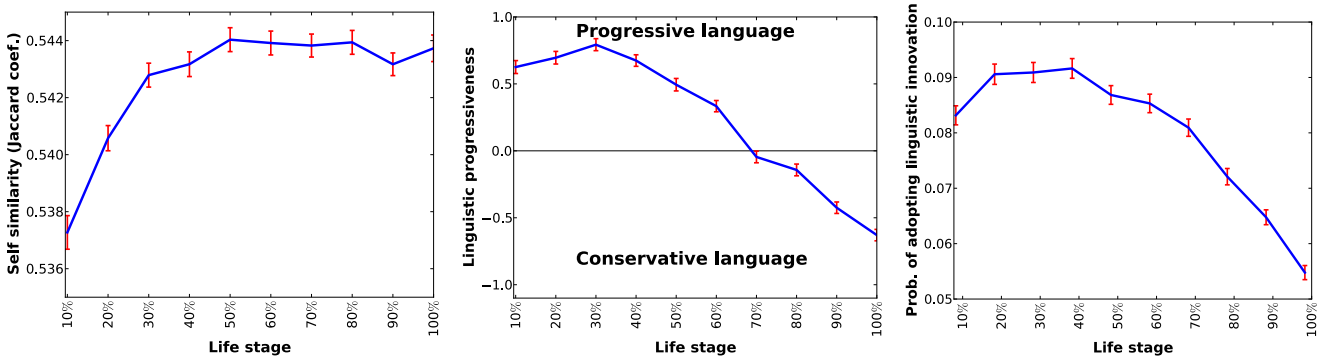
[6] More precisely, "far from the state of the community at the time the post was written": the cross-entropy of each post p is computed with respect to a community language model $\text{SLM}_{m(p)}$ contemporary with the post; this is crucially different from comparing the post with a time-invariant model of the community language since it accounts for the community-level language volatility we have discussed.

[7] We obtain the same qualitative trend for other lexical overlap measures such as cross-entropy and cosine-similarity.



(a) BeerAdvocate (b) RateBeer

Figure 6: Lifecycle: Distance from the language of the community at each life-stage, calculated as the cross-entropy of each post according to the snapshot language models of the post's month (0% is birth, 100% is death). Lower values mean "closer to the community". (a) BeerAdvocate; (b) RateBeer.



(a) Language flexibility (b) Linguistic progressiveness (c) Adoption of lexical innovations

Figure 7: Lifecycle: (a) User-language flexibility at each life-stage, computed as the Jaccard coefficient between each post and a window of 10 previous posts written by the same user. (First and last 10 reviews of each user are not represented.) Users' language rigidifies after their linguistic adolescence. (b) Linguistic progressiveness at each life-stage. Positive values indicate future-leaning language, while negative values indicate past-leaning language. (c) Probability of adopting lexical innovations at each life-stage (0% is birth, 100% is death). (BeerAdvocate; same trends hold for RateBeer.)

adolescence) and then their language rigidifies. This supports the second hypothesis: Early in their career users learn and adapt to the language of their community, but over time they stop conforming and the community slowly drifts away from them.

Users get stuck in the past. In order to gain further insight into the relation between a user's language at each life-stage and that of the community, we use the concept of *linguistic progressiveness*, which we define next. For each post p we consider the snapshot language models for the 12 months previous to the one in which p was written and the snapshot language models for the 12 months after that; we will denote with SLM_i the language model corresponding to the i-th month after (or, if i is negative, before) the month p was written. We then define the linguistic progressiveness of p as:

$$\text{Prog}(p) = \underset{-12\leq i\leq 12, i\neq 0}{\arg\min} H(p, \text{SLM}_i),$$

where $H(p, \text{SLM}_i)$ is the cross-entropy of p with respect to the SLM_i language model. Under this notation, $\text{Prog}(p) = -3$ would mean that the language of the post p appears to closest to the language used in the third month before p was written. A negative value of $\text{Prog}(p)$ indicates that p uses *conservative language* (in the sense that p uses language that looks like the language used in the past), while a positive value indicates *progressive language*. Figure 7(b) shows that users employ increasingly progressive language through their linguistic adolescence and then use language that is more and more past-leaning. This behavior has tight connections with the *adult language stability assumption* from sociolinguistics, and we will expand on these connections in Section 5.

User's reaction to lexical innovation. While our cross-entropy based measures are suitable to measure the linguistic distance from the community, the results remain opaque with respect to the actual actions through which users react to new community norms.

Type	Examples in BeerAdvocate	Examples in RateBeer
Conventions	S[mell], M[outhfeel], FLAVOR	Smell-, OVERALL, TAP
Descriptive	sandalwood, gummy, rubbery	overripe, corn-like, waxy
Other	verdict, mysterious, unexciting	nothingness, sub-par, so-so

Table 2: Examples of lexical innovations.

This motivates further investigation of a specific form of linguistic change, namely lexical innovation. As we have already discussed, the vocabulary of the community is in continuous flux with new words being constantly introduced in the community. Some of the new words will be adopted by the community and become *lexical innovation*. For the purpose of this study, a lexical innovation is a word that was never used before in the community and that was used at least 10 times by multiple users in posts discussing different products[8] over a period of 6 months after the word was first used. This way we filter out words that did not get picked up by the community and words that are user- or product-idiosyncratic. Lexical innovations include conventions (such as the ones already discussed), descriptive terms (such as fruit-words) and other types of words; using this methodology an average of 97 lexical innovations were identified each month, some of which are provided in Table 2 as examples.

We investigate the adoption of new words over the user's lifetime in Figure 7(c), which shows the probability of a user adopting lexical innovations introduced in the 3 months preceding her post, at each life-stage; we observe once again that the reaction to linguistic norms follows a lifecycle: users initially increase their rate of assimilation, which peaks at the end of their linguistic adolescence, and then follows a decreasing trend until the moment of abandonment. Figure 7(c) also shows that even though it is in the "just joined" stage when a users' language is farthest from that of the community (leftmost part of Figure 6) and when the flexibility of their language is at its height (leftmost part of Figure 7(a)), users are actually most receptive to lexical innovations (i.e., words that are not only new to the user, but also to the community) at the peak of their linguistic adolescence.

Elastic lifecycle: "All users die old". All experiments described thus far suggest the following lifecycle: After an initial period of adaptation to the language of their community—which we called linguistic adolescence in order to maintain analogy with the offline case—individuals' language patterns slowly rigidify until the moment they abandon the community. These findings confirm that the *adult language stability assumption*, central to much of the sociolinguistic work on linguistic change, holds in the online domain (see Section 5 for a detailed discussion of the implications of our findings for sociolinguistics). However, our framework also reveals a crucial difference from findings in offline settings: the moment when linguistic adolescence ends—and the user is at a peak linguistic harmony with the community—is not bound to an absolute or biological time-frame, but instead is relative to the users' own ultimate lifespan. To illustrate this we show in Figure 8 the lexical innovation lifecycle for individuals with different lifespans (i.e., different total number of contributed posts). Each curve is obtained by applying the methodology used for Figure 7(c) to sets of users with different lifespan intervals (a non-normalized x-axis is used in order to allow comparisons between the curves).

[8]Here we are even stricter, by requiring the lexical innovations to appear in reviews for products that come from different producers.

Focusing on Figure 8 brings three interesting points. First we observe that, in spite of having vastly different lifespans, users follow a similar shape in their lifecycle: an increase in the adoption of linguistic innovation followed by a decreasing trend.

Second, the moment of maximum receptiveness to lexical innovations (i.e., the end of linguistic adolescence) is not fixed (at, say, around 60 posts), but rather it is a function of the ultimate lifespan of the user. Therefore we say that the linguistic lifecycle is "elastic", in the sense that it stretches according to the ultimate lifespan of the user. All users appear to go through the same life-stages with the only difference that users that are eventually going to contribute more posts to the community end their linguistic adolescence after a larger number of reviews. A way to summarize this finding is to say that users generally die "linguistically old" (i.e., at a stage when they have relatively little reaction to linguistic change), no matter if they contribute relatively few posts to the community, or if they are heavy contributors.

Lastly, by comparing the heights of the curves in Figure 8, we can see that the level of receptivity to changing norms is also correlated with the ultimate lifespan of the user. Users that will eventually contribute more posts to the community start (and stay) at a higher level of receptivity than users that will eventually contribute less.

We also point out that the same observations regarding the elasticity of a user's lifecycle hold when we repeat the same type of split-group analysis for all other measures of linguistic change we have discussed: distance from the community, user-language flexibility, and linguistic progressiveness; they also hold in both communities.

These observations are surprising since they suggest that information about the ultimate lifespan of a user is encoded in the patterns of linguistic change exhibited in her early career; for example, if a user begins rigidifying her language after writing only 20 posts, it is unlikely that she will contribute 200 more posts before abandoning the community. In what follows we will harness these insights in the setting of a prediction task.

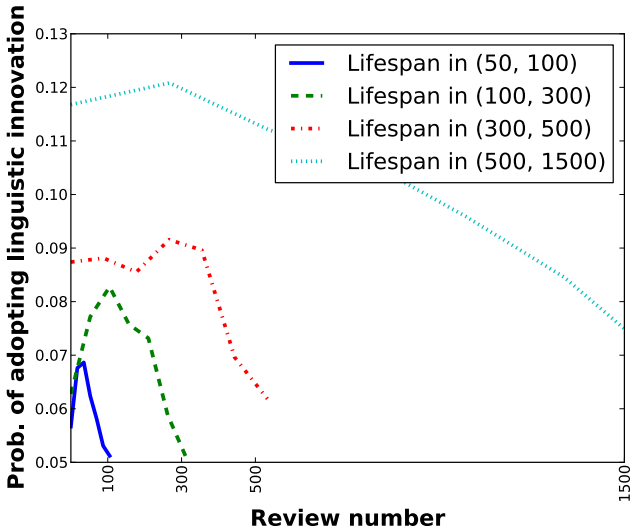


Figure 8: Lifecycle: Probability of adopting lexical innovations at each life-stage, comparing users with different lifespans (BeerAdvocate).

4. LINGIUISTIC CHANGE AS A PREDICTOR OF USER LIFESPAN

We have so far concentrated on identifying and characterizing the principles governing linguistic change in our online communities. We turn now to showing that these principles are predictive of whether a new member will soon abandon a community. Since communities thrive only if new members are able to establish themselves, this task has practical value for cultivating and maintaining vibrant online communities; the ability to identify specific groups of at-risk members early on can give community maintainers a chance to re-involve them in the community.

Our primary goal in formulating the prediction task is to use it as an example on how information about user- and community-level linguistic phenomena can inform some of the questions relevant to the maintenance of online communities. As such, the task is structured to explore relative performance gains from different types of information, rather than for optimizing raw performance per se. Overall, we hope to to contribute to a broader understanding of the process of linguistic change in online communities. In what follows, we show that features based on this new understanding lead to performance on the prediction task that improves significantly over a natural baseline. The results discussed next suggest that features based on a user's early linguistic change patterns can provide important information about that user's ultimate lifespan.

Definition of the predictive task. We define our task as predicting, for each user, whether she is among the 'departed' or the 'living'. We make these predictions based *only* on features extracted from each user's first w posts, for a small w (e.g., $w = 20$). A user is in the 'departed' class if she abandoned the community before writing m more posts for a small m (e.g., $m = 30$); we call the interval $[w, w+m]$ the *departed range*. A user is 'living' if she stayed in the community long enough to write n posts for a relatively large n (e.g., $n = 200$); we call the interval $[n, \infty]$ the *living range*.

Features used for learning. Our features are informed and designed based on the findings we reported in the previous sections. Our aim here is to illustrate the space of features that arise from these findings. To this end we consider the following five simple post-level features that we will then use to characterize a user's patterns of linguistic change:

- *Cross-entropy*: The average cross-entropy of the post according to the snapshot language model of that month. The feature draws on the characteristic U-shape pattern in users' language evolution (Figure 6).
- *Jaccard self-similarity*: The Jaccard similarity of the current post with the ten immediately preceding ones. This feature is based on the results summarized by Figure 7(a) and attempts to capture a user's flexibility to adapt to linguistic change.
- *Adoption of lexical innovations.* This feature takes value 1 if the post contains a lexical innovation introduced in community in the previous three months, and 0 otherwise. The feature is based on findings summarized by Figure 8 and aims to model a user's receptiveness towards lexical innovations.
- *First-person singular pronouns*: This feature takes value 1 if the post contains a first-person singular pronoun, and 0 otherwise. The feature is inspired by the analysis in Figure 3, where we found that, over time, users employ fewer first-person singular pronouns and by previous work suggesting that this decline can mark a sense of affiliation with a community [5, 41].
- *Number of words:* The number of words of the review. Generally, we found that long-term contributors tend to write longer reviews.

We use these post-level features to construct user-level features that capture linguistic change patterns. Because our generalizations concern the evolution of styles and linguistic norms, all of the user-level features take a common form: we divide the w observed posts into (temporally) consecutive bins of size 5 and extract for each bin the averages of each of the the five post-level features discussed above. In addition, for each of the five types of features we define a feature that indicates the bin with the maximum value[9]. Thus, for example, 'Adoption of lexical innovation' determines a family of user-level features $[\mathit{LexicalInnovation}_1, \ldots, \mathit{LexicalInnovation}_{w/5}]$ giving the average lexical innovation rate of the posts in each bin, along with another feature $\mathit{LexicalInnovation}_{\max}$ giving the index i such that $\mathit{LexicalInnovation}_i$ is maximal for this family. Our goal in using these features is to approximate the evolutionary processes we sought to characterize earlier.

As a baseline we also include two very simple yet powerful activity based family of features:

- *Frequency*: the average time between posts in each bin, as well as the index of the bin with the maximum frequency.
- *Month*: the month of the last review in each bin. These features are included to account for changes in community-wide abandonment rates (e.g., later years have a considerably higher abandonment rate, as apparent in Figure 2).

We expect these features to have especially powerful predictive value, as they directly relate to the target variable we aim to predict. In a sense, these features are much more direct ways of establishing user involvement. Moreover, prior work on user churn prediction [9, 49] found activity-based features to be by far the most powerful churn predictors. So any improvement brought by our linguistic features over this strong baseline would indicate a strong relation between a user's linguistic change patterns and her ultimate lifespan.

Experimental setup. Our models are binary logistic classifiers. Table 3 summarizes our results for the task of predicting whether a new user will stay in the community, for different choices of w (the number of reviews we observe for each user), m (defining the *departed range*) and n (defining the *living range*). The results reported are averages over 20 random train-test data splits. Each split used 70% of the data for training and 30% for testing. We treat BeerAdvocate as a 'development domain', because we used it for developing the models and experimental setting, and RateBeer as a 'test domain' in which we validate our final models on previously unseen data.

Experimental results. Table 3 compares results for 'activity-only' models with those for 'full' models (which include our five linguistic change features in addition to the activity features). We find that our linguistically-motivated features give an additional 2–12% absolute (4–40% relative) improvement in F1 score over the baseline activity-based features. Moreover, for each pairing of results (values of w), our full model improves significantly over the activity-only model, according to a paired Wilcoxon signed-rank test comparing the F1 scores from the 20 trials ($p < 0.001$).[10]

[9]For cross-entropy, the *minimum* value is selected instead, in consideration of the U-shape pattern observed in Figure 6.

[10]Given that the train and test sets are sometimes highly imbalanced, we favor evaluating these models using the F1 scores on the minor-

Community	w	Departed range	Living range	Model	Test set performance Precision	Recall	F1	Test set class sizes Departed	Living
BeerAdvocate	20	20–50	200+	Activity	77.0	41.2	53.6	327 (46%)	387 (54%)
BeerAdvocate	20	20–50	200+	Full	69.6	46.9	**56.0**	327 (46%)	387 (54%)
BeerAdvocate	40	40–100	200+	Activity	74.6	27.3	39.8	218 (36%)	378 (64%)
BeerAdvocate	40	40–100	200+	Full	66.4	31.1	**42.2**	218 (36%)	378 (64%)
RateBeer	20	20–50	200+	Activity	73.7	19.3	30.5	261 (36%)	465 (64%)
RateBeer	20	20–50	200+	Full	64.8	32.3	**42.9**	261 (36%)	465 (64%)
RateBeer	40	40–100	200+	Activity	65.9	19.6	30.0	179 (27%)	470 (73%)
RateBeer	40	40–100	200+	Full	61.3	26.3	**36.7**	179 (27%)	470 (73%)

Table 3: Predicting whether a new user is about to leave the community or will remain as an active user. The number of posts we analyze is denoted by w. The 'full' models uses all of our features, while the 'activity' models uses only activity-based features. The precision, recall, and F1 numbers given are for the target 'departing' class. For all sites and w, the full model significantly improves over the activity-only model according to a paired Wilcoxon signed rank test on the F1 scores ($p < 0.001$).

Features	F1 $w = 20$	F1 $w = 40$
Activity	30.5	30.0
+ Cross-entropy	37.4	32.2
+ Jaccard self-similarity	38.0	33.5
+ Adoption of lexical innovations	40.9	35.3
+ First-person singular pronouns	41.2	35.0
+ Number of words	42.9	36.7

Table 4: Performance improvement resulting from incrementally adding our linguistic change features to the 'activity' model (for RateBeer, our 'test community').

Feature analysis. To better understand the extend to which these improvements are explained by the exploitation of linguistic change patterns, we conduct a brief feature analysis of our full logistic models. We find that the learned coefficients are in good correspondence with the new insights brought forward in this work; for example, the 'departure class' is characterized by negative coefficients for $LexicalInnovation_1$ (i.e., a low initial rate of adoption of lexical innovation) as well as by negative coefficients for $LexicalInnovation_{\max}$ (i.e., an early end of the linguistic adolescence stage), thus corresponding to the observations summarized in Figure 8. Similar observations also hold for the cross-entropy and self-similarity features.

Furthermore, we find that each of the five families of linguistic features bring improvements in performance when added incrementally to the activity model (Table 4). At the same time, none of the features can by themselves explain the improvement reported in Table 3. This indicates that our linguistic change features complement each other in predicting a user's departure from the community.

ity class, which is also our target 'departing' class. For completeness, we note that our models also compare favorably in terms of accuracy when compared to the majority-class baseline. It should be however emphasized that the accuracy one achieves with the majority-class baseline would not translate well to a real world context, where such model could only advise a community maintainer that all or none of the community's members were leaving. In contrast, our models provide actionable intelligence in that they could help community maintainers to identify specific groups of at-risk members and try to re-involve them in the community.

5. IMPLICATIONS FOR SOCIOLINGUISTICS

The development of the framework discussed in Section 3 was guided by a large body of sociolinguistic research concerning the patterns of linguistic change in offline communities. However, we do not simply use these results passively. Rather, we show that studying very large online communities can lead to new linguistic insights and we address challenging methodological issues concerning how to track and measure change. Here we discuss how our work relates to sociolinguistic research of language change in offline communities.

In one of the earliest and most influential studies of linguistic change [22], William Labov proposed the *apparent time construct*: under the critical assumption that individuals' speech patterns are largely fixed by early adulthood, older speakers' language can be employed as a proxy for the linguistic state of the community at an earlier stage, thus providing the temporal factor necessary for studying linguistic change. This assumption, called the *adult language stability assumption,* was supported by numerous subsequent studies in a wide variety of social settings [11, 12, 23, 37, 40, 42, 47], but it is widely acknowledged that it could fail to hold in certain cases [25, 39, 46]. For instance, individuals might change their language as they age, while the community as a whole remains stable. Alternatively, a change might be simultaneously adopted by all members of the community, or just by older members. Distinguishing among scenarios like these has proved to be a fundamental challenge over the last five decades [43].

A priori, it is not at all obvious whether the adult language stability assumption suits the online world, where community-time is warped, with drastic linguistic changes arising in very short periods of time (for example, the Twitter *RT* convention achieved mainstream status in about two months [20]). Similarly, concepts like *adolescence* and *adulthood* need to be redefined to account for the vastly different interaction rates characteristic of online setting: members who interact within the community on a daily basis are likely to mature faster than members who sign on only once a month. Moreover, it has been suggested that adult language stability has biological explanations [27], and therefore is tied to actual biological aging, which is less relevant in the context of fast-paced online communities.

Our framework confirms that, in spite of these fundamental differences, the adult language stability assumption does indeed hold

in the online case: after an initial period of adaptation to the language of the community, users' linguistic patterns rigidify (Figure 7(a) in Section 3.2) and they become less likely to pick up on new community norms (Figure 7(c) in Section 3.2); the language of the "old" users is conservative and reflects the state of the community at an earlier stage (Figure 7(b) in Section 3.2).[11] However, our framework also reveals a crucial difference from the traditional formulation: the end of the linguistic adolescence is not tied to an absolute time frame, but is relative to the individual's ultimate lifespan (Figure 8 in Section 3.2). Our observations thus suggest that biological explanations are probably not the main source of adult language stability, and in general open up new avenues for the study of linguistic change in adults in highly dynamic communities.

6. ADDITIONAL RELATED WORK

In addition to the sociolinguistic and other prior work discussed above, our work draws from, and has implications for, many other research threads.

Implications for interaction in online communities. A number of early studies examined various ways in which norms arise in group interactions, including the important roles of *primacy* (the first variant of some norm tends to persist), *social status* (norms are often introduced by individuals with higher social capital) [13], minimization of *joint effort* [6], as well as mechanisms like *accommodation* [15] and *audience design* [4] by which speakers adjust their language toward that of their audience.

More recent studies have shown that similar mechanisms are at work in online communities, revealing the emergence of linguistic norms in e-mail [35], Twitter [7, 19, 20, 38, 45] and internet forums [8, 14]. This paper demonstrates the link between such community norms and the lifecycle of the individual user, showing how users are most sensitive to new norms at early stages of their career.

Our work also draws on a study that showed how new users change their language after joining a community, and demonstrated that a machine learning classifier could be trained to predict how long a user had been in a community, given linguistic features like self-introductions, references to other members, or mentions of the name of the forum [34]. Our work gives a further understanding of user change, by relating it to community change and to the entire lifespan of the user. Also, by showing that we can further predict how long a user will remain in the community, we extend the prediction task literally into the future.

Implications for the dynamics of online communities. Our work here also builds on a rich line of work studying online communities. For example, previous research has examined the dynamics of online social networks, studying the evolution of the whole network structure [21, 28], of groups inside these networks [2, 10, 16, 50], as well as the social tie creation between individual users [29]. While such studies focused on the evolution of individual components (users, groups, communities) we recognize the need to study social systems holistically. In particular, in this work we propose the study of the interaction between the evolution of individual users and that of the community at large.

[11]We also note that our results are also consistent with the standard sociolinguistic model of language change, Labov's logistic-incrementation model [24, 44], which claims that 'a general requirement of change in progress' (2001, p. 455) is the occurrence of a peak in adoption of linguistic innovation exactly at the end of linguistic adolescence—peak which we observe in Figure 7(c) and Figure 8.

There has also been a rich line of research applying methods from anthropology [48] and social psychology [30] to online communities. Issues like lurking and free-riding [36] as well as reasons for user participation in online communities [1, 26, 31] have been studied using small scale interviews and data analysis. These studies have argued that factors like group size and posting volume, newcomer status, linguistic complexity, as well as word choice, affect an individual's interaction with the community. We add an important dimension to this line of work as we employ large scale data analysis to better understand the evolution and lifecycle of individual users.

7. CONCLUSION

In this work we proposed a framework for tracking linguistic change and for understanding how individual users react to evolving community norms at different stages of their careers. By applying this framework to two large online review communities we studied the interaction between users-level and community-level evolution over a decade. This revealed that users follow a determined two-stage lifecycle: A linguistically innovative learning phase in which users align with the language of the community, followed by a conservative phase in which users stop responding to changes in community norms. We have shown that understanding patterns of linguistic change can bear practical importance for community maintainers, in that features inspired by our analysis can be used to detect early in a user's career how long she will stay active in the community.

At a higher level, the goal of this work has been to provide the foundations for reasoning about the co-evolution of users and their communities. Our work opens a range of interesting questions both in terms of sociolinguistics and of analysis of online communities. In particular, it would be interesting to investigate how the patterns of linguistic change are affected by engagement in multiple communities, as well as how users that are members of multiple communities transfer norms and conventions between communities. Furthermore, we anticipate that further analysis could potentially suggest richer linguistically and socially informed methods of identifying users that are likely to depart a community or predict a user's success in the community.

Acknowledgments We thank Julian McAuley for his assistance with the data and for his insighful suggestions, and Danqi Chen, Alex Niculescu-Mizil and the anonymous reviewers for their helpful comments. Supported in part by NSF IIS-1016909, CNS-1010921, IIS-1149837, IIS-1159679, ARO MURI, DARPA SMISC, Okawa Foundation, Docomo, Boeing, Allyes, Volkswagen, Intel, Alfred P. Sloan Fellowship and the Microsoft Faculty Fellowship.

References

[1] J. Arguello, B. Butler, E. Joyce, R. Kraut, K. Ling, C. Rosé, X. Wang. Talk to me: Foundations for successful individual-group interactions in online communities. *CHI*, 2006.

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan. Group formation in large social networks: membership, growth, and evolution. *KDD*, 2006.

[3] L. Backstrom, J. Kleinberg, L. Lee, C. Danescu-Niculescu-Mizil. Characterizing and Curating Conversation Threads: Expansion, Focus, Volume, Re-entry. *WSDM*, 2013.

[4] A. Bell. Language style as audience design. *Language in society*, 1984.

[5] C. K. Chung, J. W. Pennebaker. The psychological function of function words. K. Fiedler, editor, *Social communication: Frontiers of social psychology*. 2007.

[6] H. H. Clark, D. Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 1986.

[7] C. Danescu-Niculescu-Mizil, M. Gamon, S. Dumais. Mark my words! Linguistic style accommodation in social media. *WWW*, 2011.

[8] C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, J. Kleinberg. Echoes of power: Language effects and power differences in social interaction. *WWW*, 2012.

[9] G. Dror, D. Pelleg, O. Rokhlenko, I. Szpektor. Churn prediction in new users of Yahoo! answers. *Workshop on Community Question Answering on the Web at WWW*, 2012.

[10] N. Ducheneaut, N. Yee, E. Nickell, R. Moore. The life and death of online gaming communities: a look at guilds in world of warcraft. *CHI*, 2007.

[11] P. Eckert. Adolescent social structure and the spread of linguistic change. *Language in society*, 1988.

[12] P. Eckert. Age as a Sociolinguistic Variable. *The handbook of sociolinguistics*. Wiley-Blackwell, 1998.

[13] D. C. Feldman. The development and enforcement of group norms. *The Academy of Management Review*, 1984.

[14] M. Garley, J. Hockenmaier. Beefmoves: Dissemination, Diversity, and Dynamics of English Borrowings in a German Hip Hop Forum. *ACL*, 2012.

[15] H. Giles, J. Coupland, N. Coupland. Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge Univ Pr, 1991.

[16] S. Kairam, D. Wang, J. Leskovec. The life and death of online groups: Predicting group growth and longevity. *WSDM*, 2012.

[17] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 1987.

[18] P. Koehn. Statistical significance tests for machine translation evaluation. *EMNLP*, 2004.

[19] F. Kooti, W. A. Mason, K. P. Gummadi, M. Cha. Predicting Emerging Social Conventions in Online Social Networks. *CIKM*, 2012.

[20] F. Kooti, H. Yang, M. Cha, K. P. Gummadi, W. A. Mason. The Emergence of Conventions in Online Social Networks. *ICWSM*, 2012.

[21] R. Kumar, J. Novak, A. Tomkins. Structure and evolution of online social networks. *KDD*, 2006.

[22] W. Labov. *The Social Stratification of English in New York City*. Center for Applied Linguistics, 1966.

[23] W. Labov. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press, 1972.

[24] W. Labov. *Principles of Linguistic Change, Social Factors*. Wiley-Blackwell, 2001.

[25] W. Labov. *Principles of Linguistic Change, Cognitive and Cultural Factors*. Wiley-Blackwell, 2001.

[26] C. Lampe, E. Johnston. Follow the (slash) dot: effects of feedback on new members in an online community. *SIGGROUP*. ACM, 2005.

[27] E. H. Lenneberg. *Biological foundations of language*. Wiley, 1967.

[28] J. Leskovec, J. Kleinberg, C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2007.

[29] J. Leskovec, L. Backstrom, R. Kumar, A. Tomkins. Microscopic evolution of social networks. *KDD*, 2008.

[30] K. Ling, G. Beenen, P. Ludford, X. Wang, K. Chang, X. Li, D. Cosley, D. Frankowski, L. Terveen, A. M. Rashid, P. Resnick, R. Kraut. Using Social Psychology to Motivate Contributions to Online Communities. *Journal of Computer-Mediated Communication*, 2006.

[31] P. Ludford, D. Cosley, D. Frankowski, L. Terveen. Think different: increasing online community participation using uniqueness and group dissimilarity. *CHI*, 2004.

[32] J. McAuley, J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. *WWW*, 2013.

[33] J. McAuley, J. Leskovec, D. Jurafsky. Learning attitudes and attributes from multi-aspect reviews. *ICDM*, 2012.

[34] D. Nguyen, C. P. Rosé. Language use as a reflection of socialization in online communities. *Workshop on Language in Social Media at ACL*, 2011.

[35] T. Postmes, R. Spears, M. Lea. The formation of group norms in computer-mediated communication. *Human communication research*, 2000.

[36] J. Preece, B. Nonnecke, D. Andrews. The top five reasons for lurking: improving community experiences for everyone. *Computers in Human Behavior*, 2004.

[37] S. Romaine. *The language of children and adolescents: The acquisition of communicative competence*. Blackwell, 1984.

[38] D. M. Romero, B. Meeder, J. Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. *WWW*, 2011.

[39] G. Sankoff. Cross-sectional and longitudinal studies in sociolinguistics. *Sociolinguistics: An international handbook of the science of language and society*, 2005.

[40] G. Sankoff, H. Blondeau. Language change across the lifespan: /r/ in Montreal French. *Language*, 2007.

[41] J. C. Sherblom. Organization involvement expressed through pronoun use in computer mediated communication. *Communication Research Reports*, 2009.

[42] S. Tagliamonte. So who? Like how? Just what? Discourse markers in the conversations of Young Canadians. *Journal of Pragmatics*, 2005.

[43] S. Tagliamonte. *Variationist sociolinguistics: change, observation, interpretation*, volume 39. Wiley-Blackwell, 2012.

[44] S. Tagliamonte, A. D'Arcy. Peaks beyond phonology: Adolescence, incrementation, and language change. *Language*, 2009.

[45] O. Tsur, A. Rappoport. What's in a hashtag? Content based prediction of the spread of ideas in microblogging communities. *WSDM*, 2012.

[46] S. E. Wagner. Age Grading in Sociolinguistic Theory. *Language and Linguistics Compass*, 2012.

[47] S. E. Wagner, G. Sankoff. Age grading in the Montreal French inflected future. *Language Variation and Change*, 2011.

[48] S. M. Wilson, L. C. Peterson. The anthropology of online communities. *Annual Review of Anthropology*, 2002.

[49] J. Yang, X. Wei, M. Ackerman, L. Adamic. Activity lifespan: An analysis of user survival patterns in online knowledge sharing communities. *Proceeding of ICWSM*, 2010.

[50] E. Zheleva, H. Sharara, L. Getoor. Co-evolution of social and affiliation networks. *KDD*, 2009.