Language use as a reflection of socialization in online communities

Dong Nguyen
Carnegie Mellon University
Language Technologies Institute
Pittsburgh, PA 15213
dongn@cs.cmu.edu

Carolyn P. Rosé
Carnegie Mellon University
Language Technologies Institute
Pittsburgh, PA 15213
cprose@cs.cmu.edu

Abstract

In this paper we investigate the connection between language and community membership of long time community participants through computational modeling techniques. We report on findings from an analysis of language usage within a popular online discussion forum with participation of thousands of users spanning multiple years. We find community norms of long time participants that are characterized by forum specific jargon and a style that is highly informal and shows familiarity with specific other participants and high emotional involvement in the discussion. We also find quantitative evidence of persistent shifts in language usage towards these norms across users over the course of the first year of community participation. Our observed patterns suggests language stabilization after 8 or 9 months of participation.

1 Introduction

In this paper we use text mining and machine learning methodologies as lenses through which to understand the connection between language use and community membership in online communities. Specifically we examine an online medical support community called breastcancer.org. We present analyses of data from an active online community with the goal of uncovering the connection between language and online community membership. In particular, we will look at language changes that occur over time as people continue to participate in an online community. Consistent with the Communities of Practice theory of participation within a community (Lave and Wenger, 1991), we find increasing conformity to community norms within the first year of participation that then stabilizes as participants continue their involvement in the community.

Within the Communities of Practice view, socialization into a community begins with peripheral participation, during which individuals have the opportunity to observe community norms. Lave and Wenger's theory has been applied to both online and face-to-face communities. In an online community, observing community norms begins with lurking and reading messages before an initial post. This is termed legitimate peripheral participation, and it is during this stage that potential new members observe community norms in action. With an initial post, a user embarks upon the path of centripetal participation, as they are taking steps towards core participation.

Becoming a core member of a community means adopting community norms. Persistent language changes occur as an accumulation of local accommodation effects (Labov, 2010a; Labov, 2010b). The extent of the adoption reflects the commitment to community membership. Thus, as an individual progressively moves from the periphery of a community towards the core, their behavior will progressively grow towards conformity with these norms, although total conformity very rarely occurs. The quantitative analysis we present in the form of a regression model is consistent with this theoretical perspective and allows us to see what centripetal participation and core participation look like within the breastcancer.org community. We are able to test the robustness of these observations by using the extent

Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 76–85,
Portland, Oregon, 23 June 2011.

of conformity to community norms as a predictor of how long a member has been actively participating in an online community. We will present results from this predictive analysis as part of the quantitative evidence we provide in support of this model of community participation.

Patterns of local accommodation and of long time language change within communities have been extensively studied in the field of variationist sociolinguistics. However, with respect to online communities in particular, recent research has looked at accommodation (Danescu-Niculescu-Mizil et al., 2011; Nguyen et al., 2010) and some shorter term language changes (i.e., over a period of a few months). However, longitudinal analyses of language change spanning long time periods (i.e., more than a few months) in online communities as we present in this paper have been largely absent from the literature. Typically, long term language change in sociolinguistics requires reconstructing the past from the present using age grading techniques, since a comprehensive historical record is typically absent (Labov, 2010a; Labov, 2010b). Online communities present a unique opportunity to study long term language change from a much more comprehensive historical record of a community's development.

In the remainder of the paper, we first review prior work on computational models of accommodation and language change. We then present a qualitative view of communication within the breastcancer.org community. We then present two quantitative analyses, one that explores language change in the aggregate, and another that tests the robustness of findings from the first analysis with a regression model that allows us to predict how long a member has been active within the community. We conclude with discussion and future work.

2 Related work

For decades, research under the heading of Social Accommodation Theory (Giles et al., 1973) has attempted to layer a social interpretation on patterns of linguistic variation. This extensive line of research has provided ample quantitative evidence that people adjust their language within interactions, sometimes to build solidarity or liking, and other times to differentiate themselves from others (Eckert and Rickford, 2001).

In this line of work, people have often looked at accommodation in small discussion groups and dyadic conversation pairs. For example, Gonzales et al. (2010) analyzed style matching in small group discussions, and used it to predict cohesiveness and task performance in the groups. Scissors et al. (2009) analyzed conversational pairs playing a social dilemma game and interacting through an instant messenger. They found that certain patterns of high linguistic similarity characterize high trusting pairs. Niederhoffer and Pennebaker (2002) found linguistic style matching both at the conversation level and locally at a turn-by-turn level in dyadic conversations. Paolillo (2001) looked at the connection between linguistic variation and strong and weak ties in an Internet Relay Chat channel. Nguyen et al. (2010) found accommodation effects in an online political forum that contains discussions between people with different political viewpoints. Recently, Danescu-Niculescu-Mizil et al. (2011) showed that accommodation was also present in Twitter conversations.

Lam (2008) gives an overview of work on language socialization in online communities. We know that persistent language changes over long time periods are the accumulated result of local accommodations that occur within short-term contexts for social reasons (Labov, 2010a; Labov, 2010b). However, the process through which individuals adopt the language practices of online communities has been barely explored so far. One example of investigation within this scope is the work of Postmes et al. (2000), in which we find analysis of the formation of group norms in a computer-mediated communication setting. Specifically, they found that small groups were formed during the process and communication norms including language usage patterns were present within those groups. Over time, conformity to these norms increased. Similarly, Cassell and Tversky (2005) looked at evolution of language patterns in an online community. In this work, the participants were students from around the world participating in the Junior Summit forum '98. Cassell and Tversky found that participants converged on style, topics, goals and strategies. Analyses were computed using word frequencies of common classes (such as self references) and

Table 1: Statistics dataset.

Posts	1,562,590
Threads	68,226
Users (at least one post)	31,307
Time-span	Oct 2002 - Jan 2011

manual coding. Huffaker et al. (2006) examined a subset of the same data. When comparing consecutive weeks over a 6 week time period, they found that the language diverged. They hypothesized that this was caused by external events leading to the introduction of new words.

Our research differs from the research by Cassell and Tversky (2005), Huffaker et al. (2006) and Postmes et al. (2000) in several respects. For example, in all of this work, participants joined the community simultaneously at the inception of the community. In contrast, our community of inquiry has evolved over time, with members joining intermittently throughout the history of the community. Additionally, our analysis spans much more time, specifically 2 years of data rather than 3 or 4 months. Thus, this research addresses a different question from the way community norms are first established at the inception of a community. In contrast, what we investigate is how new users are socialized into an existing community in which norms have already been established prior to their arrival.

We are not the first researchers to study our community of inquiry (Jha and Elhadad, 2010). However, prior work on data from this forum was focused on predicting the cancer stage of a patient rather than issues related to language change that we investigate.

3 Data description

We analyze one of the largest breast cancer forums on the web (http://community.breastcancer.org/). All posts and user profiles of the forum were crawled in January 2011.

The forum serves as a platform for many different kinds of interactions, and serving the needs of a variety of types of users. For example, a large proportion of users only join to ask some medical questions, and therefore do not stay active long. In fact, we find that a lot of users (12,349) only post in the first week after their registration. The distribution of number of weeks between a user's last post and registration date follows a power law. However, besides these short-term users, we also find a large number of users who appear to be looking for more social involvement and continue to participate for years, even after their disease is in remission.

This distinction in types of users is reflected in the forum structure. The forum is well organized, containing over 60 subforums targeting different topics. Besides specific subforums targeting medical topics (such as '*Stage I and II Breast Cancer*' and '*Radiation Therapy - Before, During and After*'), there are subforums for certain population groups (such as '*Canadian Breast Cancer Survivors*' and '*Singles with breast cancer*'), for social purposes (such as '*Growing our Friendships After Treatment*', '*Get Togethers*', and '*CyberSisters Photo Album*') and non cancer related purposes (such as '*Humor and Games*'). In many of the subforums there are specific threads that foster the formation of small sub communities, for example threads for people who started chemotherapy in a certain month.

In the data we find community norms of long time participants that are characterized by forum specific jargon and a style that is highly informal and shows familiarity with specific other participants and high emotional involvement in the discussion. We infer that the forum specific jargon is distinct from what we would find in those users outside of it, in that that there are places in the forum explaining commonly used abbreviations to new users. We also observe posts within threads where users ask about certain abbreviations used in previous posts. Some of these abbreviations are cancer related and also used in places other than the forum, such as dx (diagnosis), and $rads$ (radiation, radiotherapy). Thus, they may be reflective of identification with a broader community of cancer patients who are internet users. Other often used abbreviations are dh (dear husband), dd (dear daughter), etc. We also observed that users frequently refer to members of the community by name and even as $sister(s)$.

Now let us look at some examples illustrating these patterns of language change. We take as an example a specific long-time user. We start with a post from early in her participation, specifically from a couple of days after her registration:

I am also new to the forum, but not new to bc, diagnosed last yr, [..] My follow-up with surgeon for reports is not until 8/9 over a week later. My husband too is so wonderful, only married a yr in May, 1 month before bc diagnosed, I could not get through this if it weren't for him, never misses an appointment, [...] I wish everyone well. We will all survive.

The next two posts[1] are from the same user, 2 to 4 years after her registration date. Both posts are directed to other forum members, very informal, and contain a lot of abbreviations (e.g. '*DH*' (Dear Husband), '*DD*' (Dear Daughter), '*SIL*' (Son in Law)).

Gee Ann I think we may have shared the same 'moment in time' boy I am getting paid back big time for my fun in the sun. Well Rose enjoy your last day of freedom - LOL. Have lots of fun with DH 'The Harley'. Ride long and hard (either one you choose - OOPS).

Oh Kim- sorry you have so much going on - and an idiot DH on top of it all. [..] Steph- vent away - that sucks - [..] XOXOXOXOXOXOXOX [..], quiet weekend kids went to DD's & SIL on Friday evening, they take them to school [..], made an AM pop in as I am supposed to, SIL is an idiot but then you all know that.

This anecdotal evidence illustrates the linguistic shift we will now provide quantitative evidence for.

4 Patterns of language change

4.1 Approach

In this section we aggregate data across long time participants and look at global patterns of language change. Specifically, we will analyze patterns of change in the first year after registration of these members, and show how language patterns consistently become more different from the first week of participation and more similar to the stable pattern found within the second year of data. Furthermore, when comparing consecutive weeks we find that the difference increases and then stabilizes by the end of the first year. The unit of analysis is one week of data. Because there are multiple ways to measure the similarity or difference between two distributions, we explore the use of two different methods. The first metric we use is the Kullback-Leibler (KL) divergence. Larger values indicate bigger differences in distribution. P represents the true distribution. Note that this metric is asymmetric.

$$KL(P,Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

We also explore using the Spearman's Rank Correlation Coefficient (SRCC), which measures the similarity of two rankings:

$$SRCC = 1 - \frac{6 \sum_i d_i^2}{n(n^2-1)}$$

Where d_i is the difference between the ranks of word i in the two rankings and n is the total number of words.

4.2 Sampling

In this analysis, we begin by aligning the data of every member by registration date. We then aggregate posts of all users by week. Thus, in week 1, we have the posts from all users during the first week after their registration. Note that the actual week in time would not be the same for each of these users since they did not all register at the same time. In this way, a week worth of data represents the way users talk after the corresponding number of weeks after registering with the community rather than representing a specific period of time. Because our dataset spans a large time period of time (i.e. more than 8 years), it is very unlikely that patterns we find in the data reflect external events from any specific time period.

As discussed before, a large proportion of members only post in their first week after registration. These short time members might already initially differ from members who tend to participate longer in the forum. Therefore, it might confuse the model if we take these short time members into account. We may observe apparent changes in language that are artifacts of the difference in distribution of users across weeks. Thus, because we are interested in language change specifically, we only consider posts of long-term participants.

[1] Names are replaced in example

In addition, we have limited our focus to the initial two-year period of participation, because it is for this length of participation that we have enough users and enough posts to make a computational model feasible. We have also limited ourselves to examining high frequency words, because we have a large vocabulary but only a limited amount of data per week. Two weeks can look artificially similar if they both have a lot of non-occurring words. In summary, taking above considerations into account, we applied the following procedure:

- We only look at the first 2 years, for which we still have a large amount of data for every week.
- We only look at members who are long-term participants (2 years or longer), this leaves us with 3,012 users.
- For every week, we randomly sample an equal number of posts (i.e., 600 from each week). All posts are taken into account (i.e. both responses as well as thread openings).
- We only look at the distribution change of high frequency words (words occurring at least 1,000 times), this leaves us with 1,540 unique words. No stemming or stop word removal was done.

4.3 Comparison with early and late distributions

Using the dataset described in the previous section, we compare the language of each week during the first year after registration with language in the very first week and with language in the second year.

First we analyze whether language in the first year becomes more similar to language used by members in their second year as time progresses. We therefore compare the word distributions of the weeks of the first year with the overall word distribution of the second year. We apply KL divergence where we consider the distribution of the second year as the 'true distribution'. The result is shown in Figure 1. We see that the KL divergence decreases, which means that as time progresses, the word distributions look more like the distribution of the second year. Fitting a Least Squares (LS) model, we get an intercept of 0.121033 and slope of -0.001080

Figure 1: KL divergence between weeks in first year and overall second year.

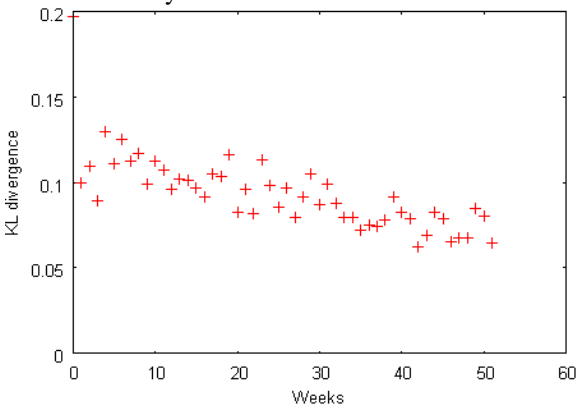
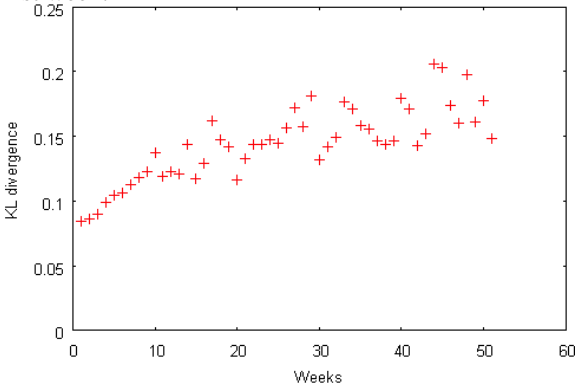


Figure 2: KL divergence between weeks in first year and first week.



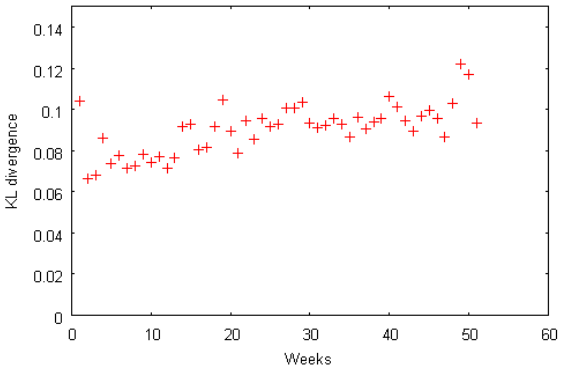
($r^2 = 0.5528$). Using the Spearman Rank Correlation (SRCC) and fitting a LS model, we observe the same pattern ($r^2 = 0.6435$).

Our second analysis involves comparing the distributions of the first year (excluding the first week), with the distribution of the first week. The result is shown in Figure 2. We see that the KL divergence increases, meaning that as time progresses, the word distributions become less similar with the first week. (KL: $r^2 = 0.6643$, SRCC: $r^2 = 0.7962$).

4.4 Comparing consecutive distributions

We now compare the distributions of consecutive weeks to see how much language change occurs in different time periods. For KL divergence we use the symmetric version. Results are presented in Figure 3 and show a divergence pattern throughout the first year that stabilizes towards the end of that first year of participation. (KL: $r^2 = 0.4726$, SRCC: $r^2 =$

Figure 3: KL divergence between consecutive weeks.



0.8178). The divergence pattern was also observed by Huffaker et al. (2006) (related, but not equivalent setting, as mentioned in the literature review). We hypothesize that the divergence occurs because users tend to talk about a progressively broader set of topics as they become more involved in the community. To confirm this hypothesis, we compare the distributions of each week with the uniform distribution. We indeed find that as time progresses, the distributions for each week become more uniform. (KL: $r^2 = 0.3283$, SRCC: $r^2 = 0.6435$).

5 Predicting membership duration

In the previous section we found strong patterns of language change in our data. We are interested in the extent to which we can automatically predict *how many weeks* the user has been a member, using only text or meta-features from that specific week. Identifying which features predict how long a member has been active can give more detailed insight into the social language that characterizes the community. In addition, it tells us how prominent the pattern is among other sources of language variation.

5.1 Dataset

For this analysis, we set up the data slightly differently. Now, rather than combine data across users, we keep the data from each user for each week separate so we can make a separate prediction for each user during each week of their participation. Thus, for each person, we aggregate all posts per week. We only consider weeks in the first two years after the registration in which there were at least 10 posts with at least 10 tokens from that user.

Table 2: Statistics dataset.

	#Docs	**#Persons**	**#Posts**
Training	13,273	1,591	380,143
Development	4,617	548	122,489
Test	4,571	548	134,141

5.2 Approach

Given an input vector $\mathbf{x} \in \mathbb{R}^m$ containing the features, we aim to find a prediction $\hat{y} \in \mathbb{R}$ for the number of weeks the person has been a member of the community $y \in \mathbb{R}$ using a linear regression model: $\hat{y} = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$ where β_0 and $\boldsymbol{\beta}$ are the parameters to estimate. Usually, the parameters are learned by minimizing the sum of squared errors.

In order to strive for a model with high explanatory value, we use Linear Regression, with L1 regularization (Tibshirani, 1996). This minimizes the sum of squared errors, but in addition adds a penalty term $\lambda \sum_{j=1}^{m} |\beta_j|$, the sum of absolute values of the coefficients. λ is a constant and can be found by optimizing over the development data. As a result, this method delivers sparse models. We use Orthant-Wise Limited-memory Quasi-Newton Optimizer (Andrew and Gao, 2007) as our optimization method. This method has proven to establish competitive performances with other optimization methods, while producing sparse models (Gao et al., 2007).

Because our observations suggest that language change decreases as members have been active longer, we also experimented with applying a log transformation on the number of weeks.

5.3 Features

For all features, we only use information that has been available for that particular week. We explore different types of features related to the qualitative differences in language we discussed in Section 3: textual, behavioral, subforum and meta-features.

5.3.1 Textual features

We explore the following textual features:

- *Unigrams* and *bigrams*.
- *Part of Speech* (POS) bigrams. Text was tagged using the Stanford POS tagger (Toutanova et al., 2003).

- *LIWC* (Pennebaker et al., 2001), a word counting program that captures word classes and stylistic features.

- *Usernames*. Because some of the usernames are common words, we only consider usernames of users active in the same thread.

- *Proper names*. We obtained a list containing common female names. We ranked them according to frequency in our dataset, and manually deleted common words in our dataset, such as *happy, hope, tuesday* and *may*, from our list.

- *Slang words*. We manually compile a list of common abbreviations and their whole words counterpart. We then count the number of abbreviations and the number of whole words used in the post. The feature value then is $(\#abbrev - \#wholewords)/\#totalwords$. Because in some contexts no abbreviations can be used, this feature takes into account if the user actually chose to use the abbreviation/whole word, or if there was no need for it.

No stemming or stopword removal is used. Frequencies are normalized by length.

5.3.2 Behavioral features

We also explore additional features that indicate the behavior of the user:

- Ratio (posts starting threads) / (total number of posts).

- Number of posts.

5.3.3 Subforum features

We include as features the distribution of subforums the member has posted in. This captures two intuitions. First, it is an approximation of the current phase in the cancer process for that member. For example, we noticed that most of the new users have just been diagnosed, while long term users have already finished treatment. Because the subforums are very specific (such as '*Not Diagnosed with a Recurrence or Metastases but Concerned*'), we expect these features to give a good approximation of the phase the user is currently in. In addition, these subforums also give an indication of the user's interest. For example, whether the user posts mostly in medical forums, or mostly in the social orientated subforums.

Table 3: Results reported with Pearsons correlation (r).

Run	# Features	Raw (r)	Log (r)
Unigrams + Bigrams	43,126	0.547	0.621
POS	1,258	0.409	0.437
LIWC	88	0.494	0.492
Proper names	1	0.185	0.186
Usernames	1	0.150	0.102
Slang	1	0.092	0.176
Behavior	2	0.139	0.243
Subforum	65	0.404	0.419
All above	44,542	0.581	0.649
All above + Person	46,133	0.586	0.656

5.3.4 Other features

Most of the persons appear multiple times in our dataset (e.g. multiple weeks). To help the model control for idiosyncratic features of individual users, we include for every person a dummy variable associated with that user's unique identity. This helps the model at training time to separate variance in language usage across users from general effects related to length of participation. Note that we do not use these features as test time.

5.4 Results

We experimented with individual types of features as well as all of them aggregated. The results (correlations) can be found in Table 3. The features having the most weight for long time participants in our best model (All incl. Person, Log) are presented in Table 4. We see that for most features the performance was higher when applying the log transformation. This was especially the case with the unigrams and bigrams features. For some features the difference was less, such as for proper names and the subforum features. This could indicate that these features have a more linear pattern as time progresses, while word patterns such as unigrams tend to stabilize earlier. We find that both stylistic patterns (such as POS) as well as patterns indicating conformity (social behavior, slang words) are individually already very predictive.

In our best performing model, we find that both

Table 5: Qualitative grouping of textual features.

Type	Short time members	Long time members
Abbreviations	Husband	My DD (Dear Daughter), Your PS (Plastic Surgeon)
Social networks		Facebook, fb
Greetings	Hi all	Hi girls, Hi gals
I versus other	LIWC-I, My, Me	LIWC-other, We, Sisters
Social support		Hugs, Condolences, So sorry
Thanking	Thanks, Thanx, Thx	
Forum		Bc org, On bco
Introducing	Newbie, New here, Am new	
Asking information	Info, LIWC-qmarks	

Table 4: Top 10 features of long term users.

Feature	Weight
META - slang	0.058362195
META -propername	0.052984915
year	0.050872918
META - [person1]	0.050708718
META - [person2]	0.040548104
months	0.040400583
META - [person3]	0.039806096
LIWC - Othref	0.036080545
META - [person4]	0.035605996
POS - nnp prp	0.035033650

the slang and proper name features get a high weight for long time participants. Furthermore, we observe that a lot of the person meta features are included in the model when it is trained, although as mentioned we do not use these features at testing time. The fact that the model assigns them weight indicates that idiosyncratic features of users explain a lot of variance in the data. Our best performing model has 3,518 non zero features. In Table 5 we qualitatively grouped and contrasted features that were more associated with short-term or long-term members. We see that long-term members show much more social behavior and familiarity with each other. This is shown to references to each other, more social support, references to social networks and ways of greeting. They furthermore talk about the forum itself more often by using the abbreviation '*bco*'. Short term members are characterized by words that are used when they introduce themselves to others.

Thus we find that long time participants are characterized by informal language, containing many forum specific jargon, as well as showing emotional involvement with other forum members. Our best run obtained a correlation of $r = 0.656$, giving an r^2 value of 0.430. This means that 0.43 of the variation can be explained by our model. Since there are many other factors that influence the writing of users, it is understandable that our model does not explain all the variance.

6 Discussion

As discussed widely in previous literature, people become socialized into communities over time through their interactions with community members. The extent of conformity to group norms reflects commitment to the group. Our first study showed evidence of increasing conformity to community norms through changes in simple word distributions. The second study then tested the robustness of these findings through a prediction task and extended the language features of the first study.

Since community members tend to conform increasingly to community norms over time, although the target class for our predictive model is time, it is reasonable to assume that what the model really learns to predict is how long average community members have been around by the time they sound like that. In other words, one can think about its time prediction as a measure of how long it sounds like that person has been in the community. The model would therefore overpredict for members who move from the periphery to the core of a community faster than average while underpredicting for those who do so more gradually. This would be consistent with the

idea that rate of commitment making and conformity is person specific.

There are two limitations that need to be addressed regarding the present studies. First, there are certain factors that influence the rate of adoption to the forum that we are not able to take into account. For example, some people might have already been reading the forum for a while, before they actually decide to join the community. These people are already exposed to the community practices, and therefore might already show more conformity in the beginning than others.

Second, our experiments involved one online community targeting a very specific topic. Due to the nature of the topic, most of the active users come from a small subpopulation (mostly women between 40-60 years). Therefore, it is a question how well these results can be applied to other online communities.

As a future application, a model that can capture these changes could be used in research related to commitment in online communities.

7 Conclusion

It is widely accepted that persistent language change in individuals occurs over time as a result of the accumulation of local processes of accommodation. Although previous research has looked at accommodation within short periods of time, including recent research on social media data, persistent language change as a result of longer term involvement in an online community is still an understudied area.

In this paper we have presented research aiming to close this gap. We have analyzed data from a large online breast cancer forum. Analyzing data of long time members, we found strong patterns indicating language changes as these members participated in the community, especially over the course of their first year of participation.

We then presented a regression approach to predict how long a person has been a member of the community. Long time participants were characterized by showing more social behavior. Furthermore, they used more forum specific language, such as certain abbreviations and ways of greeting. Due to the nature of our dataset, language was also influenced by external factors such as changes in the cancer process of individuals.

Although our observations are intuitive and agree with observations in previous, related literature regarding socialization in communities, it is still a question whether our observations generalize to other online communities.

In our current work we have looked at changes across users and across contexts. However, it is well known that individuals adapt their language depending on local interactions. Thus, a next step would be to model the process by which local accommodation accumulates and results in long term language change.

Acknowledgments

The authors would like to thank Michael Heilman for the regression code and Noah Smith for ideas for the regression experiments. This work was funded by NSF grant IIS-0968485.

References

Galen Andrew and Jianfeng Gao. 2007. Scalable training of l1-regularized log-linear models. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 33–40, New York, NY, USA. ACM.

Justine Cassell and Dona Tversky. 2005. The language of online intercultural community formation. *Journal of Computer-Mediated Communication*, 10:16–33.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words! linguistic style accommodation in social media. In *Proceedings of WWW*.

Penelope Eckert and John R. Rickford. 2001. *Style and Sociolinguistic Variation*. Cambridge: University of Cambridge Press.

Jianfeng Gao, Galen Andrew, Mark Johnson, and Kristina Toutanova. 2007. A comparative study of parameter estimation methods for statistical natural language processing. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 824–831, Prague, Czech Republic, June. Association for Computational Linguistics.

Howard Giles, Donald M. Taylor, and Richard Bourhis. 1973. Towards a theory of interpersonal accommodation through language: some canadian data. *Language in Society*, 2(02):177–192.

Amy L. Gonzales, Jeffrey T. Hancock, and James W. Pennebaker. 2010. Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37(1):3–19, February.

David Huffaker, Joseph Jorgensen, Francisco Iacobelli, Paul Tepper, and Justine Cassell. 2006. Computational measures for language similarity across time in online communities. In *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech*, ACTS, pages 15–22, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mukund Jha and Noémie Elhadad. 2010. Cancer stage prediction based on patient online discourse. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, BioNLP '10, pages 64–71, Stroudsburg, PA, USA. Association for Computational Linguistics.

William Labov. 2010a. *Principles of Linguistic Change, Volume I, Internal Factors*. Wiley-Blackwell.

William Labov. 2010b. *Principles of Linguistic Change, Volume I, Social Factors*. Wiley-Blackwell.

Wan S. E. Lam. 2008. Language socialization in online communities. In Nancy H. Hornberger, editor, *Encyclopedia of Language and Education*, pages 2859–2869. Springer US.

Jean Lave and Etienne Wenger. 1991. *Situated Learning. Legitimate peripheral participation*. Cambridge: University of Cambridge Press.

Dong Nguyen, Elijah Mayfield, and Carolyn P. Rose. 2010. An analysis of perspectives in interactive settings. In *Proceedings of the 2010 KDD Workshop on Social Media Analytics*.

Kate G. Niederhoffer and James W. Pennebaker. 2002. Linguistic style matching in social interaction.

John C. Paolillo. 2001. Language variation on internet relay chat: A social network approach. *Journal of Sociolinguistics*, 5:180–213.

James W. Pennebaker, Roger J. Booth, and Martha E. Francis, 2001. *Linguistic Inquiry and Word Count (LIWC): A Computerized Text Analysis Program*.

Tom Postmes, Russell Spears, and Martin Lea. 2000. The formation of group norms in computer-mediated communication. *Human Communication Research*, 26(3):341–371.

Lauren E. Scissors, Alastair J. Gill, Kathleen Geraghty, and Darren Gergle. 2009. In cmc we trust: the role of similarity. In *Proceedings of the 27th international conference on Human factors in computing systems*, CHI '09, pages 527–536, New York, NY, USA. ACM.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.