Level of domain expertise impacts language use

Benjamin Peloquin
Stanford University bpeloqui@stanford.edu

Abstract

How does the amount we know about a topic impact the language we use to describe it? To gain traction on this question, we collect a data set of 50 thousand RateBeer.com reviews and operationalize *level of experience* as number of reviews written by individual users. Within this framework, we re-examine evidence from previous studies which suggests that experience level influences language use. We explore this evidence in two ways – first as a statistical language analysis and second as a classification task. Our findings suggest that level of experience reliably predicts linguistic features in our review data and that we can classify users' experience level from linguistic data alone.

1 Introduction

How does the amount we know about a topic impact the language we use to describe it? And, given the way someone talks about a topic, can we recover the amount they know about it? Previous studies indicate that the accumulation of domain specific experiences can influence the way we talk about those experiences. These findings have been leveraged to create better product recommendation systems (McAuley & Leskovec, 2013), to study individual- and group-level linguistic change (Danescu-Niculescu-Mizil et al. 2013) and in expert identification in Question Answer communities (Pal et al. 2011).

These previous studies, however, employed different operational definitions of *expertise* or *experience level*, and evidence for linguistic-level differences was only studied indirectly through feature-engineering for machine learning tasks or through secondary, post-hoc analyses.

In the current study, we use a data set of over 50 thousand reviews from RateBeer.com. Unlike the previous studies cited above, we operationalize *experience level* straightforwardly – as the number of reviews written by individual users.

We first orient the reader to previous work which has examined the impact of experience-level on language use. In each study, we highlight specific hypotheses directly or indirectly indicated by findings from the research. In an effort to unify these findings, we use *number of reviews written* by users to operationalize *level of experience*. Using this definition, we conduct a language analysis, re-examining findings from the previous studies. To preview our results, we find that more experienced users are more likely to use fewer first person pronouns (FPPs), less likely to use negation, display more expansive vocabulary and employ more formulaic grammatical structures. Following our language analysis, we consider the task of classifying users' level of experience based on language data alone, comparing four classifiers to a baseline chance model.

1.1 McAuley & Leskovec (2013)

In McAuley & Leskovec (2013), the authors investigated the hypothesis that successful product recommendations should take into account not only user *tastes*, but also *level of experience*. Under this formulation, some products may be more accessible to users than others, depending on how experienced the user is at a given point in time. Similarly, they proposed that user tastes evolve over time and that "the very act of consuming products will cause users' taste to change and evolve."

The authors assessed four models which formalized different modes of maturity evolution in users' tastes, using a data set of 15 million product reviews, including reviews from RateBeer.com (data used in the current study).

They began with a standard latent factor recommender system which predicted ratings of user/item pairs (u, i) based on:

$$rec(u, i) = \alpha + \beta_u + \beta_i + \langle \gamma_u, \gamma_i \rangle$$

where α is a global offset, β_u and β_i are user and item biases and γ_u and γ_i are latent user and item features. They extended this basic model to accommodate temporal (experience) information by fitting a latent variable e_{ui}, which represented the the experience of a user u at time t_{ui}. To model the evolution of experience over time, they constrained each user's experience level to be a non-decreasing function of time - so that as a user rated more products, their experience-level always increased. They did so using a monotonicity constraint on time and experience:

$$\forall u, i, j \quad t_{ui} \geq t_{uj} \Rightarrow e_{ui} \geq e_{uj}$$

Experience level was modeled as a categorical variable that could take on discrete values $e_{ui} \in \{1...E\}$. (In their reported results they constrained the domain of values for $e \in \{1, ..., 5\}$.) The final recommender system was implemented as function of this experience parameter e_{ui} such that:

$$rec(u, i) = rec_{ui}(u, i)$$
$$= \alpha(e_{ui}) + \beta_u(e_{ui}) + \langle \gamma_u(e_{ui}) + \gamma_i(e_{ui}) \rangle$$

Under their formulation, a user became more experienced (or stayed at the same experience level) as they rated additional products. The latent factor, e, enabled the authors to model a user as "progressing" between recommender systems as they gained experience. Critically, the notion of *expertise* in this study is an **interpretation of this latent parameter**, e.

In total, the author's examined four models, each encoding a different hypothesis about user experience progression - (1) as a group at uniform intervals, (2) as individuals at uniform intervals, (3) as a group at learned intervals and (4) as individuals at learned intervals.

Following a series of model comparisons, the authors conducted a "qualitative analysis," exploring the impact of experience on review ratings using the model's latent experience parameter, e, as a proxy for level of experience. In their analysis, *experts* were defined as users who reached the highest latent factor experience value ($e = 5$) and novices the lowest value ($e = 1$). Given this set-up they made the following observations.

1. Experts' ratings are more predictable (lowest MSE) by their model.

2. Experts tend to have cohesive ratings - they agree with each other more than other groups.

3. Experts have more extreme views - experts give "better" products higher ratings and "worse" products lower ratings.

4. There are cohesive genres of products (beer styles in RateBeer.com) that experts tend to like or novices tend to like.

We take these findings as indirect evidence that experts and novices may display distinct behavioral patterns as it pertains to review writing and ratings. While *level of experience* in this study is an interpretation of the model's latent parameter, e, the nature of the parameter does fit an intuitive notion of expertise (i.e. something that is accumulated over time, through experience in a particular domain).

In our current work, we examine findings suggested by this study, using the total number of reviews written by a given user as a proxy for *level of experience*. In particular, we can explore the hypothesis suggested by the findings – (1) the most experienced users' ratings were the most predictable and (2) more experienced users' ratings tended to cohere – by examining whether *more experienced users display more formulaic (predictable) language in their review writing*.

1.2 Danescu-Niculescu-Mizil et al. (2013)

Danescu-Niculescu-Mizil et al. (2013) studied linguistic change using RateBeer.com and BeerAdvocate.com review data. They explored this phenomenon in two ways - first as an analysis of the relationship between user-life cycle and language and, second, as a prediction task, in which the authors attempted to predict a user's lifetime within a community using linguistic features from their first analysis.

Users showed a measurable life-cycle with respect to specific linguistic aspects. In particular, new users tended to display higher degrees of receptiveness to linguistic norms of the community up to around a third of the their eventual lifespan in the community. As users became more experienced they also became increasingly unreceptive to changing linguistic norms until the moment they abandoned the site. The authors showed that the basic form of this trend was robust across user lifespans (e.g. it held for users who used the site for shorter and longer periods).

The authors examined two types of linguistic change: 1) at the user level and 2) at the community level. At the user level, the authors found that *use of first person pronouns (FPPs) decreased* with experience. Conversely, use of *beer specific vocabulary* increased with experience. Additionally, they found that users employed increasingly progressive language through their linguistic adolescence and then used language that was more past-leaning. They tied this finding to the sociolinguistic adult language stability assumption.

At the community level, the authors noted two interesting conventions – the gradual adoption of "smell" over "aroma" and the introduction of more "fruity" words over the course of the community life-cycle. Interestingly, the authors also found that community language entropy across months declined as the community matured, so the community as a whole became more "cohesive" or predictable across the nearly 7-year time-frame considered.

Using insights from their initial analysis, the authors attempted to predict whether a user will have "departed" or still belong to the community given that users first w posts (for some small w such as $w = 20$). A user was classified as "departed" if she abandoned the community before writing m more posts for a small m (e.g. $m = 30$). They used the following features in a binary classification task: (1) *Cross-entropy for the post according to a snap shot language model of that month*, (2) *Jaccard self-similarity of the current post with ten immediately preceding posts*, (3) *adoption of lexical innovations*, and (4) *review length*.

As a baseline, the authors fit a logistic regression model with only two features, *frequency* (the average time between posts) and *month* (the month of the last review). They found significant model improvement incorporating the five linguistic features on both precision and recall. Best model performance ($F1 = 56.0$) occured with a full model, $w = 20$ and the departed range $20 - 50$.

While Danescu-Niculescu-Mizil et al. (2013) did not focus on *expertise* directly, they observed interesting linguistic features that differed based on a user's *level of experience* within their individual life-cycle. In our current work, we can examine these observations directly – instead of assessing *within-user* impacts of experience we will explore a *between-user* analysis. In particular, we can investigate the observations: (1) *use of FPPs decreases with experience*, (2) *length of reviews increases with experience*, (3) *internal cohesion of review language increases with experience*, (4) *vocabulary increases with experience*.

1.3 Pal et al. (2011)

While the two previous studies approached expertise indirectly, either as an interpretation of a model parameter or relative to an individual user's life-span, Pal et al. (2011) explicitly set out to build a machine learning (ML) classifier to identify experts in a Question Answer community.

Question answering communities provide information to users who participate in social interactions with one another. In Pal et al. (2011) the authors examined the TurboTax Live Community (TLLC), a community that allows users to ask and answer tax related questions.

Given the importance of expert contributions in communities like TLLC, companies invest in methods for identifying and cultivating these members, often giving them special status once they have been identified. At the time of this project, Intuit employed a team to manually identify experts – a high precision, low recall, and overall time consuming enterprise. The primary objective in Pal et al. (2011) was to use ML techniques to automatically identify high-potential users in the first few weeks of participation.

The authors identified two primary dimensions for feature engineering, which the called "motivation" and "ability."

Motivation encoded the idea that an expert should be highly motivated to help others and *ability* captured the idea that an expert should have the ability to answer questions correctly. A set of linguistic features were included under *ability* which captured the *manner* in which users answered questions, with the hypothesis that experts should display higher rates of politeness and clarity. The used the following features: (1) *presence of spelling mistakes*, (2) *use of profanity*, (3) *use of FPPs* and (4) *use of negation*.

The authors compared a SVM classifier and C4.5, a Decision Tree classifier, using 10-fold cross-validation. Overall the authors found a fairly large discrepancy in terms of precision and recall between these two methods, with C4.5 performing better on recall and getting a better F1 score overall (0.37 for SVM and 0.5 for C4.5).

Having identified a set of users from their classifiers, the authors asked Intuit employees experienced with the task of expert identification to evaluate the classifications. About three quarters of the users identified by the ML classifiers were confirmed to have the potential to become super-users (experts).

While Pal et al. (2011) were interested in leveraging all the information available for the expertise classification task, in our current study we are most interested in the linguistic features they considered. We can interpret this feature engineering as implicitly encoding hypotheses about the impact of expertise on language use. In particular, we will examine: (1) *use of FPPs*, (2) *use of profanity*, (3) *occurrence of spelling mistakes*, and (4) *occurrence of negation*.

2 Language analysis

2.1 Current dataset

Table 1 includes a description of the data collected for our current work. While both McAuley & Leskovec (2013) and Danescu et al. (2013) used RateBeer.com data, they used different operational definitions of *level of experience*. In our own data collection, we adopted a user-level sampling strategy – generating a set of user IDs and then crawling both user home-pages, which contained information such as the user's total number of reviews written, geographic location and number of followers, and up to 50 of their most recent reviews written. We imposed this upper limit of 50 reviews per user to avoid biasing our data set with reviews written only by users with many reviews. This user-centric sampling strategy ensured that we had sufficient power to examine hypotheses across multiple user-experience levels.

RateBeer.com data	
Number of reviews	52,751
Number of users	4,939
Median review length	35
Users with only 1 post	1,879
Users with more than 500 posts	204

Table 1: Description of data collected for the current study

2.2 Empirical framework

In the following section, we examine a set of hypotheses proposed by the previous work. In each study the authors found distinct features they believed were defining of experts or were tied to an individual user's experience level.

McAuley & Leskovec (2013) operationalized *experience* as a latent factor in a recommender system and found that users with more experience tended to have more predictable and similar ratings. We investigate a secondary hypothesis suggested by this finding - that more experienced users will display more formulaic (predictable) language in their reviews, which we measure through a *self-similarity* metric.

Danescu-Niculescu-Mizil et al. (2013) assessed the way user language changed based on time in a community. The found that as level of experience increased, users tended to: use fewer first person pronouns (FPPs), write longer reviews, use more beer-specific vocabulary (which we measure via *linguistic diversity*) and tended to "rigidify" linguistically (e.g. become more similar, which we measure via *self-similarity*).

In Pal et al. (2011), the authors built an expertise classifier which included linguistic level features such as: occurrence of spelling mistakes, use of profanity, use of FPPs and use of negation.

	Hypothesis	**Authors**	**Dir**	**Result**
1	FPPs	D+P	↓	↓*
2	lex-div	D	↑	↑*
3	negation	P	UNK	↓*
4	profanity	P	↓	Null
5	length	D	↑	↑*
6	similarity	D+JL	↑	↑*
7	spelling	P	↓	Null

Table 2: Hypotheses based on findings from previous studies. The **Hypothesis** column contains hypotheses investigated. **Authors** column contains the studies the hypotheses were derived from (*ML* - McAuley & Leskovec, 2013; *D* - Danescu-Niculescu-Mizil et al. 2013; *P* - Pal et al. 2011). **Dir** contains the *predicted* direction of the effects. **Result** column shows model coefficient directions and corresponding p-values (* indicates $p < 0.001$ and "Null" indicates a non-significant effect size).

To examine the hypotheses enumerated in Table 2, we fit a series of mixed effects models, regressing the dependent variable of interest against log10(*number of reviews*), *number of friends*, *review length*, *review overall score*, *review taste score*, *review aroma score*, *review palate score*, *review global score* and *user-* and *beer-level random effects*, which was the maximal structure that converged. We used the R packages *lme4* for model fitting and *lmertest* to report corresponding p-values.

2.3 First person pronoun usage

First person pronoun (FPPs) use decreased as users gained experience in Danescu-Niculescu-Mizil (2013) and Pal et al. (2011) included this as a linguistic feature in their classifiers. We identified FPP usage in reviews using a simple lexicon and look-up function in R. Results indicate that *number of user ratings* is a significant predictor of *FPPs* ($\beta = -0.27, t = -9.98, p < 0.001$), such that an increase in the number of ratings (user experience) predicts fewer FPPs.

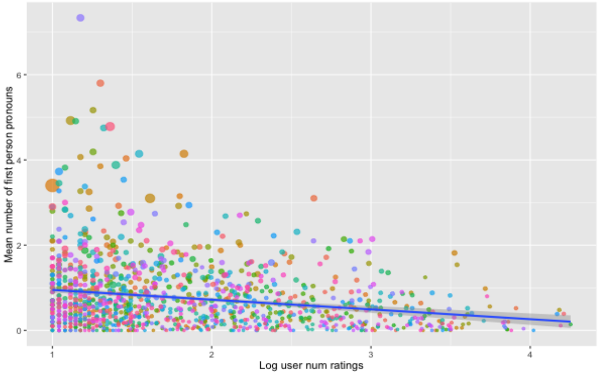

Figure 1: Use of FPPs as a function of experience level. Horizontal axis plots user experience level. Vertical axis plots mean usage of FPPs. Each data point represents the average for a single user and the point's diameter corresponds to variance of the estimate.

2.4 Lexical diversity

Danescu-Niculescu-Mizil et al. (2013) observed increased use of specialized vocabulary as users gained experience. We use *lexical diversity* as a proxy for this finding. *Lexical diversity* measures how many different words are used in a text usually formalized through the ratio of *types* (number of distinct words) to *tokens* (total number of words). We use Carroll's *corrected type-token ratio* which is a ratio of types T to tokens N in a corpus, corrected for number of tokens:

$$\text{CTTR} = \frac{T}{\sqrt{2N}}$$

There are number of type-token ratios (TTRs) that measure lexical diversity, however Carroll's CTTR is considered standard (Benoit *quanteda* R package). Results indicate that *lexical diversity* increases with experience ($\beta = 0.00285, t = 7.05, p < 0.001$). We will return to this finding after commenting on *linguistic self-similarity*.

2.5 Negation

Pal et al. (2011) included a feature for presence of negation in their classification models. They made no prediction about the anticipated direction of the effect. Previous studies have shown that use of negation in review data increases when reviewers have bad experiences (Jurafsky et al. 2014) and is generally used more as review ratings decrease (Potts, 2011). As an aside, we regressed *overall rating* on the same predictor set from above and saw evidence for these previous findings – use of negation increases as *overall rating score* decreases ($\beta = -0.0015, t = -7.9, p < 0.001$). We also see evidence that use of negation *decreases* with experience ($\beta = -0.0062, t = -8.0, p < 0.001$). We assessed negation in reviews using a lexicon of negation cues and a simple look-up function in R.

2.6 Profanity usage

Pal et al. (2011) included a feature for occurrence of *profanity* encoding the hypothesis that more experienced users will display higher levels of politeness. To identify profane language we created a lexicon of the top 18 scoring (for offensiveness, familiarity and personal usage) words from Janschewitz (2011), a study on taboo words and used a simple look-up function in R. Results indicate that there is no effect of experience on profanity usage ($\beta = -0.00003, t = -0.025, p = 0.98$).

2.7 Review length

Danescu-Niculescu-Mizil et al. (2013) observed increases in review length as users became more experienced. We measured review length by extracting the number of tokens in each review. Results confirm the Danescu-Niculescu-Mizil et al. (2013) observation as *review length* increased as a function of experience level ($\beta = 2.80, t = 5.54, p < 0.001$).

2.8 Linguist self-similarity

Dinescu-Niculescu-Mizil et al. (2013) and McAuley & Leskovec (2013) both observed that reviews and ratings appeared to cohere more as users became more experienced. We formalized this qualitative observation through a metric we have dubbed *linguistic self-similarity*. Linguistic self-similarity (self-sim) is computed by taking the average cosine similarity of all the pairwise combinations for a user's u reviews $r \in \{1...n\}$ for a user with n reviews:

$$\text{self-sim}_u = \frac{\sum_{i=1}^{n} \sum_{j=1}^{i-1} cos(\text{review}_i, \text{review}_j)}{\frac{n*(n-1)}{2}}$$

and $cos(r_1, r_2)$ is calculated by taking the dot product for two reviews represented as token count vectors and dividing by the product of their euclidean lengths:

$$cos(r_1, r_2) = \frac{\vec{r_1} \cdot \vec{r_2}}{|\vec{r_1}||\vec{r_2}|}$$

A user with a higher *self-similarity* score means that on average their reviews tend to use similar language (cohere as in the sense of Danescu-Niculescu-Mizil, 2013 and McAuly & Leskovec, 2013).

Because *self-similarity* is an average score for each user, instead of fitting a mixed-effects model we fit a multiple regression model, regressing *self-similarity* score on *number of ratings*, *average review length*, *average overall score*, *average taste score*, *average aroma score*, *average appearance score*, *average palate score*, *average beer global score* and *number of different styles tried by user*.

Results indicate that *self-similarity* increases as a function of experience ($\beta = 0.05, t = 7.75, p < 0.001$). This is an interesting result, especially in light of the finding that *lexical diversity* also increases as a function of experience. We might expect *lexical diversity* and *self similarity* to be negatively correlated - intuitively if your reviews have greater lexical diversity shouldn't they also be more distinct? An informal analysis of some of the users with the most reviews indicates that the review process becomes highly systematized, almost formulaic, such that more experienced users employ "fill in the blank," style structures, largely reusing grammatical themes between reviews. Simultaneously, these more experienced users display more expansive vocabularies (leading to higher lexical diversity) which they use to "fill in the blanks."

2.9 Spelling errors

Pal et al. (2011) included a feature for occurrence of *spelling mistakes* encoding the hypothesis that more experienced users will display greater "clarity" in their language. We identified spelling mistakes using the R package *qdap*. Results indicate that there is no effect of experience on number of spelling mistakes, ($\beta = 0.105, t = 0.825, p = 0.41$).

2.10 Discussion

Operationalizing *level of experience* through the number of reviews written, we assessed seven hypotheses motivated by previous studies. We found evidence for a number of these hypotheses, such as the observations that as users gain experience their reviews tend to *contain fewer FPPs*, *employ more diverse vocabulary*, *contain less negation*, *become longer* and *become more formulaic*. We return to these findings, grounding them in a qualitative analysis of randomly selected user reviews, in the general discussion.

3 Classifying experience-level

Our language analysis provides evidence that more experienced users in the RateBeer.com community display distinct differences in the language of their reviews. We now turn to the task of recovering a user's experience level from linguist data alone. We treat this problem as a classification task and examine a series of machine learning classifiers.

To proceed, we bin all users into experience quartiles. The breaks are *class 1* (less than 18 reviews), *class 2* (between 18 and 72 reviews), *class 3* (between 72 and 276 reviews) and *class 4* (more than 276 reviews).

3.1 Classifiers

We compared four classifiers against a baseline model, which returned the most common class (~ 0.25). We used two standard machine learning classifiers - Naive Bayes (NB) and Random Forest (RF). Both the NB and RF classifiers were given tri-, bi- and unigram features with limited text normalization (case-folding). We conducted minimal hyper-parameter tuning with the RF classifier due to run-time constraints, eventually choosing a model with 100 trees.

We also explored a *language model* (LM) approach to classification. These classifiers consisted of four language models (either unigram with Laplace add-one smoothing or trigram Stupid-backoff models) trained on data from each quartile. Classifications were made by returning the class label of the LM which assigned the lowest overall perplexity score to a given review.

Mean accuracy for all models was assessed via 10-fold cross-validation on a subset of our total data (30,000 reviews). Figure 2 plots mean accuracy ratings with 95% confidence intervals for our classifiers.

3.2 Results and discussion

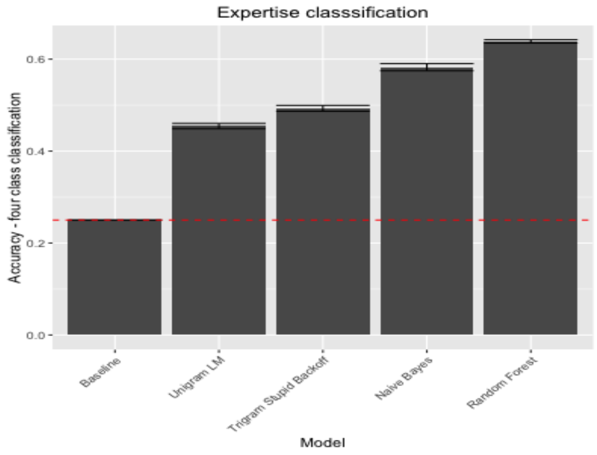

Figure 2: Mean accuracy scores for classifiers surveyed in the current work

We compared four classifiers against a baseline, measuring accuracy on a four-class classification task (see Figure 2). Model comparisons were made using 10-fold cross-validation. All models performed substantially better than a chance, baseline model, which simply returned the majority class (since reviews were binned in quartiles baseline accuracy was 0.25). The RF classifier performed best, achieving accuracy of 0.64. NB performed second best with an accuracy of 0.58. While the language model classifiers were outperformed by RF and NB, results for both were substantially above chance. The trigram Stupid-backoff classifier nearly achieved fifty percent accuracy (0.49). These language models can be interpreted as a form of supervised generative models, similar to Naive Bayes. Future studies may want to experiment with more sophisticate smoothing techniques including linear interpolated language models or Kneser-Ney smoothing.

3.3 General discussion

In the current study, we examined how the amount of reviews written by a user in the RateBeer.com community impacted the language of their reviews. We approached this topic by performing a language analysis of RateBeer.com data, using seven hypotheses indirectly or directly indicated by previous work. Table 2 contains our findings.

We now ground these findings in a qualitative analysis of two users with different experience levels (22 vs 15,292 reviews written), first comparing their reviews for the same beer and then examining additional review each has written.

1. User 245351 (22 reviews), Beer "10-barrel-joe": *Tried this at the 10 Barrel pub on International IPA Day. Very nice aroma that has a nice sweetness and hop notes. Mouthfeel is excellent and the flavor is well balanced. It is very hoppy but isn't overwhelming. Lots of citrus flavors and very drinkable. I'm drinking my too quickly. Excuse me, waitress. May I please have another?*

2. User 96974 (15292 reviews), Beer "10-barrel-joe": *On tap @ 10 Barrel. Nice orange colored beer with a sticky white head. Tons of grapefruit, orange, and peach. Pretty big on the pine as well. Mild grass and bready malt. Flavor comes off a bit herbal but has those nice citric fruit qualities to it as well. Slight sweetness of bread, melon, and stone fruit. Really quite enjoyed this beer.*

In our quantitative analysis we observed declines in use of FPPs and increases in lexical diversity as a function of experience. These appear to correspond to review narratives oriented around precise descriptions of the beer's perceptual characteristics. By contrast, less experienced reviewers appear to include more beer-irrelevant information, such as descriptions of the occasion or information about the mindset or personality of the reviewer. Clearly, we see this at work in User 245351's first review (e.g. "...I'm drinking my too quickly. Excuse me, waitress. May I please have another").

Likewise, instead of using generic language such as "Lots of citrus flavors," User 96974 lists the actual ingredients detected (e.g. "Tons of grapefruit, orange and peach... Slight sweetness of bread, melon, and stone fruit"). The observed decreases in use of negation from our quantitative analysis may also emerge from this orientation around perceptual acuity – more experienced users strive to describe the beer exactly as it is, rather than what it is not. We see some evidence for this in the flavor descriptions above. User 245351 writes, "It is very hoppy but isn't overwhelming" in describing the flavor while User 96974 writes, "Flavor comes off a bit herbal but has those nice citric fruit qualities to it as well."

Generally, more experienced users appear to be more verbose. This coincides with their focus on "exact" descriptions and also leads to longer overall reviews. However, we also see more experience users adopt regular content and grammatical forms. For example, here are two more reviews from the same two users as above, for different beers.

1. User 245351 (22 reviews), Beer "BridgePort Smooth Ryed": *If you are hop friendly then this is a drinkable and refreshing beer! Overall, I think Bridgeport has done a nice job with this seasonal and it has been my go-to beer this Spring!*

2. User 96974 (15292 reviews), Beer "Tahoe Mountain Festivus": *On tap @ Diving Dog. Dark brown in color with a khaki head. Some roasted malt and a lot of dark sugars–candi sugar, molasses, and toffee. Some interesting spice notes like mint, pepper, anise, and cinnamon. Has a bit of a tart plum and raisin note on the finish. Medium bodied, some creamy carbonation, and a bit of bitterness. Not really my thing.*

User 96974 begins again with the method of drinking (tap vs bottle) and moves into a description of the characteristics, providing approximately a sentence each to appearance, flavor, palate, aroma, and a last qualitative assessment. By contrast, User 245351's second review has little correspondence with their first. There is little description of the actual beer and appears highly subjective (e.g. "I think Bridgeport has done a nice job..."). These observations correspond to our quantitative findings of increased self-similarity between reviews as users become more experienced.

Taken together, findings from our quantitative analysis and informal qualitative analysis indicate that more experienced users strive for apparent *objectivity* and *professionalism* in their reviews. As reviewers become more experienced their reviews become increasingly formulaic and simultaneously more descriptive and detailed.

As a secondary test of the findings from our language analysis we attempted to classify a user's experience level, given some review, using only language data from that review. If experience-level differences (including some we may have overlooked) were truly characteristic of different reviews, then our classifiers should be sensitive to those differences and we should expect accuracy beyond a chance model. Binning our users into experience quartiles we trained two standard machine learning classifiers (Naive Bayes and Random Forest) and explored two "Language model classifiers" using a unigram with Laplace add-one smoothing and a trigram with Stupid-backoff smoothing language models. We observed two primary findings. First, we were able to gain significant traction on this multi-class classification task across all our models. Every model, including the worst performing unigram language model, performed significantly better than baseline. The second finding was the that our exploratory language models, while nearly achieving 50% accuracy, both performed worse than NB and RF classifiers with tri-, bi- and unigram features. RF achieved the best score with 0.64 accuracy.

Previous studies have provided indirect evidence for the impact of user experience on language use. In the current study, we consolidated findings from these studies and introduced a novel, unifying approach to re-examine them directly. Adopting *number of reviews* as a proxy for *level of experience*, we found significant differences in language based on users' experience level. We confirmed these findings using a classification task which attempted to classify users based solely on the language of their reviews. These results represent valuable information both for the socio-linguistics community concerned with the study of language differences between groups as well as organizations who might benefit from the identification of expert users in their communities.

Acknowledgments

Thanks to Chris Potts and Will Monroe for valuable discussion on the paper and Bill McCartney for his advice throughout the project.

References

Benoit K., Nulty P. 2016. *quanteda: Quantitative Analysis of Textual Data*. R package version 0.0.6-9. https://CRAN.R-project.org/package=quanteda

Danescu-Niculescu-Mizil C., West R., Jurafsky D., Leskovec J., Potts C. 2013. *No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities*. WWW 2013.

Janschewitz K. 2008. *Taboo, emotionally valenced, and emotionally neutral word norms*. Behavior Research Methods Vol 40, Issue 4, pp 1065-1077.

Jurafsky D., Chahuneau V.,Routledge, B., & Smith N. 2014. *Narrative framing of consumer sentiment in online restaurant reviews*. First Monday, 19(4)

McAuley J. & Leskovec J. 2013. *Frmom Amateurs to Connosoisseurs: Modeling the Evolution of User Expertise through Online REviews*. WWW 2013.

Pal A., Frazan R., Kraut R. 2011. *Early Detection of Potential Experts in Question Answering Communities*. CMU Research Showcase; Human-Computer Interaction Institute.

Potts C. 2011. *On the negativity of negation*. SALT 20. 636-659.