Discovering User Attribute Stylistic Differences via Paraphrasing

Daniel Preoţiuc-Pietro, Wei Xu and **Lyle Ungar**
Computer & Information Science
University of Pennsylvania
danielpr@sas.upenn.edu, {xwe,ungar}@cis.upenn.edu

Abstract

User attribute prediction from social media text has proven successful and useful for downstream tasks. In previous studies, differences in user trait language use have been limited primarily to the presence or absence of words that indicate topical preferences. In this study, we aim to find linguistic style distinctions across three different user attributes: gender, age and occupational class. By combining paraphrases with a simple yet effective method, we capture a wide set of stylistic differences that are exempt from topic bias. We show their predictive power in user profiling, conformity with human perception and psycholinguistic hypotheses, and potential use in generating natural language tailored to specific user traits.

Introduction

The large volume of social media data presents us with a unique opportunity to learn about the users authoring the texts. Many user traits are known to impact the content of a user's posts. Conversely, using only the posted text, we can automatically predict a number of traits using machine learning methods. These include age (Rao et al. 2010), gender (Bamman, Eisenstein, and Schnoebelen 2014), location (Eisenstein et al. 2010), personality (Schwartz et al. 2013), political affiliation (Volkova, Coppersmith, and Van Durme 2014) or impact (Lampos et al. 2014). Modelling author traits has proven useful for tasks such as text classification (Hovy 2015) and sentiment analysis (Volkova, Wilson, and Yarowsky 2013).

Users choose at posting time both the topic of the post and the style in which it is expressed. Topical choice is heavily influenced by author attributes. In previous analyses, most predictive and distinctive text features show primarily topical differences (Van Durme 2012; Bergsma and Van Durme 2013; Vogel and Jurafsky 2012). Stereotypically, males post more about sports, while women are post more about fashion. Features like 'wife' and 'beard' are most predictive of male users, while 'husband' and 'purse' are predictive of female users (Sap et al. 2014). In an analysis of occupational class (Preoţiuc-Pietro, Lampos, and Aletras 2015), the authors show that users with higher skilled jobs post more about politics and finance, while lower skilled users post more about personal things like beauty and leisure. They also showed that lower skilled users use more elongations, hinting at stylistic differences.

The style in which users write is much more subtle. In this study, we isolate stylistic differences by using paraphrase pairs and clusters. Paraphrases represent alternative ways to convey the same information (Barzilay 2003), using either single words or phrases. By studying occurrences within these paraphrase pairs and clusters, we directly present the difference of stylistic lexical choice between different user groups, while minimizing the conflation of topical differences. Example paraphrase clusters for three user traits (gender, age and occupational class) are presented in Table 1 where: lower occupational class is associated with higher usage of more colloquial word choice ('gramps', 'grandpa'); females use longer and more nuanced words ('wonderfully', 'delightfully') instead of standard words ('fine', 'well'); older users seem to use more formal and complex words to express congratulations ('commended'). Table 2 shows an example of the same cluster of word paraphrases across all three traits. In this example, high occupational class users use much more formal and elaborate word choice ('gratifying', 'enjoyable'). Older users prefer similar words to high occupational class users, only with a slightly different preference order (Kendall's $\tau = 0.51$). Finally, for this example, the gender paraphrase ranking is almost orthogonal to the others ($\tau = 0.06$ gender $-$ age, $\tau = 0.2$ occupation $-$ gender), showing that these occupy a separate stylistic spectrum.

We analyze the three user attributes – gender, age and occupational class – each using a large Twitter corpus. We first assess the predictive performance of user attributes using paraphrase pairs, differentiating between topical and stylistic influences. Results show significant stylistic differences between user traits and these still hold significant predictive power, albeit to a lesser extent than topical content. We then measure how well our method captures meaningful stylistic differences between all author traits by comparing it with human judgements, showing that our scores match with human perception. Further, we explore psycholinguistic theories about stylistic differences between groups. Our methods show that females choose words which are perceived as being happier, but use less syllables. Higher occupational class and more senior users prefer words which are longer, have more syllables, and prefer more abstract words. This work is useful

Gender	wonderfully > delightfully > beautifully > fine > well > good > nicely > superbly
Age	applauded >commended >salute >praised >pay tribute >congratulated
Occupational Class	grandfather > granddad > grandpa > grandad > gramps

Table 1: Example paraphrase clusters ordered by preference. Female, high occupational class and $\geqslant 30$ years old are to the left.

Female	$\geqslant$ 30 y.o.	High Occ. Class
charming (-0.58)	delightful (-0.96)	gratifying (-0.33)
delightful (-0.49)	splendid (-0.55)	enjoyable (-0.30)
gratifying (0.03)	enjoyable (-0.46)	delightful (-0.17)
splendid (0.07)	gratifying (-0.04)	pleasant (-0.06)
good (0.12)	pleasant (0.01)	charming (-0.04)
pleasing (0.18)	charming (0.03)	splendid (-0.01)
nice (0.24)	pleasurable (0.13)	good (0.08)
pleasurable (0.24)	nice (0.53)	nice (0.17)
pleasant (0.33)	good (0.53)	pleasing (0.21)
enjoyable (0.38)	pleasing (0.80)	pleasurable (0.77)
Male	< 25 y.o.	Low Occ. Class

Table 2: Example paraphrase cluster ordered by gender, age and occupational class. The number in brackets represents the magnitude with which the word is related to each trait.

for tailoring content to users, for example in dialogue systems or targeted advertising. All paraphrase user trait scores described in this study are openly available.[1]

Data

We use three Twitter data sets to extract phrase statistics. All data is tokenised (Preoţiuc-Pietro et al. 2012), filtered for English using the *langid.py* tool (Lui and Baldwin 2012) and part-of-speech tagged using the Twitter model of the Stanford tagger (Derczynski et al. 2013).

Gender

We use a data set consisting of 104,500,740 posts from 67,337 Twitter users (31,682 males and 35,655 females). The gold gender labels are obtained from the users' self-reported information in their linked accounts on other networks such as Facebook or MySpace, a method used in (Burger et al. 2011; Volkova, Wilson, and Yarowsky 2013).

Occupational Class

This data set consists of tweets from 5,191 Twitter users labeled with one of the nine major occupational classes as given by the UK Standard Occupational Classification (Preoţiuc-Pietro et al. 2015). The users are spread across the nine major classes, with the skill level required for an increasing as the class number is lower. The data set consists of 10,796,836 tweets and was introduced in (Preoţiuc-Pietro, Lampos, and Aletras 2015). We consider high occupational class (or higher skilled) users to be those belonging to classes 1-2 (2,292 users) and low occupational class (or lower skilled) users from classes 5-9 (1,617 users).

[1]http://figshare.com/articles/Paraphrase_choice_based_on_user_traits/1613525

Age

The age data set consists of 5,091 users mapped to their real age. Their age is identified by mining posts of the type 'Happy X birthday to @USER' (Volkova and Bachrach 2015). We divide the users in two groups: $\geqslant 30$ years old (958 users) and < 25 years old (2,907 users). The final data set consists of 690,077 tweets.

Inducing Stylistic Paraphrases

We use the Paraphrase Database (PPDB) (Ganitkevitch, Van Durme, and Chris Callison-Burch 2013) as our source of paraphrases. PPDB 2.0 (Pavlick et al. 2015) contains over 100 million paraphrases derived from a large collection of bilingual texts by pivoting methods. Each paraphrase pair comes with ranking scores about its quality and reliance based on a supervised regression model using distributional similarity and other features. We use paraphrase pairs that have an equivalence probability of at least 0.2.

Given a paraphrase pair, we use phrase (here, 1-3 grams) occurrence statistics computed over our data sets to measure the phrase choice difference over user attributes.To score which user attribute (we exemplify with gender in this paragraph) favors a phrase w, we compute the scores Male(w) and Female(w) which represent the average normalized male and female, respectively, usage of phrase w. For each phrase we compute a score:

$$\text{Gender}(w) = \log\left(\frac{\text{Female}(w)}{\text{Male}(w)}\right) \quad (1)$$

Within a paraphrase pair (w_1, w_2), the difference $\text{Gender}(w_1) - \text{Gender}(w_2)$ measures the stylistic distance between user trait groups. We apply the same method for both occupational class and age. Our methodology is similar to the work of (Pavlick and Nenkova 2015) who studied paraphrasing in the context of formality and complexity.

We also use paraphrases clusters that are created by using the transitive closure of pairwise paraphrases (the paraphrase relationship in PPDB is not symmetric). Within these clusters, we subtract the mean phrase scores. This adjusts for topic prevalence e.g., males swear *a priori* more than females, so scores in a swear-word cluster would be skewed towards males without this adjustment. This leads to a score of 0 representing a reference point across all clusters.

Analysing Inter-Trait Differences

We use the three data sets to perform an exploratory study on trait similarities in terms of topic and style, the latter measured using paraphrases. We first measure the correlations between all phrases using the score defined in the previous

Traits		Correlation
Age (< 25 y.o.)	Occ.Class (Low)	0.366
Occ.Class (Low)	Gender (Female)	0.355
Age (< 25 y.o.)	Gender (Female)	0.269

Table 3: Pearson correlations between traits computed using scores over all vocabulary phrases. Positive label is mentioned in brackets.

Traits		Correlation
Occ.Class	Gender	0.241
Age	Occ.Class	0.178
Age	Gender	0.162

Table 4: Pearson correlations between paraphrase cluster standard deviations. This highlights which traits show similar divergent styles.

section. Results comparing all pairs of the three datasets are presented in Table 3.

We observe that by this direct comparison using all vocabulary phrases, scores are most correlated between users under 25 years old and users with low occupational class. This is perhaps not surprising, as younger users are more likely to have a job with a lower skill level. The next highest correlation is intriguingly between low occupational class and female gender. Lastly, female gender word scores are also correlated with younger users to a lesser extent.

However, we turn to paraphrases and paraphrase clusters in order to measure stylistic language choice and test if the same patterns hold as by using all vocabulary phrases. As exemplified in Table 2, for each trait a paraphrase cluster consists of a phrase ranking in terms of preference and the scores reflect the divergence in usage. Within each cluster, a higher standard deviation thus indicates a higher divergence in usage between the two user opposing groups. The highest average cluster standard deviation is obtained for age ($\mu = 0.684$), with occupational class ($\mu = 0.466$) and gender ($\mu = 0.258$) considerably lower. We thus see that the highest difference in word use is observed between age groups, perhaps partially influenced by our choice of age groups.

We further analyze the correlations between the clusters ranked by standard deviation in order to see which groups of paraphrases have similar divergence. Results are shown in Table 4.

Occupational class and gender paraphrase clusters have the most divergent clusters in common, while age and gender have the least, similarly to the analysis over all phrases. However, the most relevant aspect for phrase choice, is the order in phrase preference within a paraphrase cluster. For example, in Table 2 the paraphrase cluster is ordered very similarly by high occupational class and by older users (Kendall's $\tau = 0.51$). In this case, gender is almost orthogonal to the other two dimensions ($\tau = 0.06$ gender $-$ age, $\tau = 0.2$ occupation $-$ gender). To analyze this in more detail, for each paraphrase cluster we compute a measure of rank similarity between all pairs of user traits. The average rank similarity across all clusters is presented in Table 5.

Traits		Correlation
Age (< 25 y.o.)	Occ.Class (Low)	0.318
Age (< 25 y.o.)	Gender (Female)	0.306
Occ.Class (Low)	Gender (Female)	0.305

Table 5: Average Kendall τ rank correlation between paraphrase cluster usage compared acrosst different user traits. Spearman rank correlation ρ and Pearson correlation reveal similar patterns.

The results show, consistently to Table 3, that younger age and lower occupational class share the most overlap in terms of paraphrase phrase choice. The correlation between the phrase choices of younger users and females is almost as high as that between low occupational class users and females. This is despite that on overall phrase usage, the correlation between low occupational class and females was considerably higher. The above analysis demonstrates that stylistic paraphrase choice reveals different patterns to those extracted by analyzing all phrases.

Predicting User Attributes

In this experiment, we test the predictive power of the set of words part of paraphrases in comparison to all vocabulary phrases. We randomly select 80% of the users to build probabilities for each phrase and keep 20% for use in evaluating prediction accuracy. We use the Naïve Bayes classifier to assign a score to each user. As we evaluate our models using ROC AUC and correlation, we do not need to estimate a prior class distribution based on the training data. This results in the Naïve Bayes likelihood for each user and task being equal to the dot product between the log probability of the word belonging to one class (similar to the measure we previously defined) and the user phrase distribution vector.

To measure the influence of paraphrase choice, we compare the performance of the full model using all phrase scores (with a relative frequency of over 10^{-5} in each dataset) to the model using only phrases appearing in at least one paraphrase pair (except for trivial paraphrases which differ only in stopwords) and the rest of the phrases separately. The former is a proxy for stylistic choice while the latter is a proxy for topical information. Results of evaluating the user gender score against the true gender are shown in Table 6 in ROC AUC (Table 6a) and point-biserial correlation (Table 6b). All correlations are significant ($p < 10^{-10}$, two tailed t-test).

We notice some general patterns which hold across both metrics. Overall, the phrases also part of paraphrase pairs lead to lower performance when compared to the phrases not part of paraphrases. However, these still hold significant prediction performance over chance, showing that they contain important cues. For age, adding the paraphrase features improves the prediction performance significantly ($\sim .04$ ROC AUC). Not coincidentally, age was the feature that showed the highest paraphrase differences in the analysis from the previous section. Adding paraphrase features to gender prediction leads to similar results. In the case of occupational

	Gender	Age	Occ.Class
Random Baseline	0.5	0.5	0.5
Only Paraphrases	0.691	0.742	0.700
Phrases w/o Paraphrases	0.784	0.861	0.870
All phrases	0.765	0.901	0.795

(a) Area under the Receiver Operating Characteristic Curve (ROC AUC)

	Gender	Age	Occ.Class
Random Baseline	0	0	0
Only Paraphrases	0.321	0.354	0.271
Phrases w/o Paraphrases	0.449	0.513	0.474
All phrases	0.442	0.579	0.419

(b) Point-biserial Correlation

Table 6: User attribute prediction results. Using only paraphrases that capture more stylistic rather than topical differences between different user demographic groups, our method still show good predictive power comparing to using all phrase (1-3 grams) features.

class predictions, actually adding these features hurts predictive performance. With respect to author trait prediction performance, age is also easiest to classify (.901 ROC AUC), followed by occupational class (.870 ROC AUC) and gender (.784 ROC AUC).

Human Perception of Writing Style

We perform a crowdsourcing experiment in order to validate our measure of stylistic difference within paraphrase pairs and check if this perceptible to human raters. We use the paraphrase scores as a indicator of stylistic divergence and for each user trait, we present a random set of 100 paraphrase pairs with a difference larger than 0.2 to the workers to rate.

Each paraphrase pair was rated by 5 workers on Amazon Mechanical Turk (MTurk). The workers were asked to pick which phrase they think is more likely used by a female/higher skilled/older user rather than a male/lower skilled/younger user. To eliminate possible biases, we randomized the order of the two phrases in a paraphrase pair when presenting them to different workers.

Overall, in around 70% of the pairs (see Table 7), the majority vote of human annotators agrees with the automatic measures described in the previous sections. This demonstrates that differences in word choice are perceptible by humans and can be traced to different user traits. The highest agreement was with respect to age (73.3%), again highlighting that age carries the largest stylistic differences. Similarly to the prediction results, occupational class differences were the lowest, although still significantly above chance. Table 7 shows some examples of paraphrases correctly rated by crowdsourcing workers. Table 8 shows some representative examples where our automatic rating is different from human judgements. Many disagreements are caused by word polysemy (e.g. ‘classes’) or shortening (e.g. ‘yrs’) or rating out-of-context (e.g. ‘legend’). The inter-annotator agreements between crowdsourcing workers are on the low side but not out of norm for capturing the subtle linguistic style difference: Fleiss’ Kappa $\kappa = 0.110, 0.402, 0.242$ for gender, occupational class and age respectively. Further selection of crowdsourcing workers may improve the rating reliability (Gao et al. 2015), which we leave for future work.

Agreement Rate (%)	Examples
Gender (68.5%)	husband > hubby fascinating > charming dame > lady comfortable > comfy laugh > giggle
Occupational Class (67.2%)	unnecessary > useless many > alot decline > fall unveils > presents proposals > suggestions
Age (73.7%)	chaos > mess impressive > amazing victory > winning assault > attack excellent > great

Table 7: Agreement between human ratings and automatic measures based on statistics over text. The words on the left are more likely to be used by male/higher-skilled/older people.

Disagreement Rate (%)	Examples (human rating)
Gender (25.8%)	okay > nice handmade > homemade fierce > brutal
Occupational Class (30.3%)	humour > wit story > legend indie > independent
Age (24.0%)	classes > groups years > yrs protective > defensive

Table 8: Disagreement between human ratings and automatic measures based on statistics over text. The words on the left are more likely to be used by male/higher-skilled/older people according to human raters. A small proportion of word pairs ($\approx$ 5%) received ambiguous human judgements and are excluded from Table 7 and 8.

Linguistic Hypotheses

We investigate a number of psycholinguistic hypotheses about phrase choice and style by using our paraphrase based method. We argue the phrase choice within a paraphrase pair excludes the topical influence that confounds studies using all words (Sarawgi, Gajulapalli, and Choi 2011). Using unigram paraphrases we study if any user group is more likely to use a word based on the following properties:

Word length We compute the difference in word length in a paraphrase pair as a simple proxy for word complexity.

Number of syllables We compute the difference in the number of syllables in a paraphrase pair as another proxy for word complexity.

Feature	Gender	Age	Occ.Class
Word length	.089**	.158**	.211**
#Syllables	.047**	.077**	.110**
Word rareness	-.053**	-.028**	-.034**
Happiness	-.051**	-.022*	-.026*
Concreteness	-.048**	-.037**	-.124**

Table 9: Correlation coefficients between paraphrase pair word differences and user group differences in usage. The **positive** class is **male** (for gender), **higher skill level** (for occupational class) and users $\geqslant 30$ **years old** (for age). $p < 0.05^{*}$, $p < 0.001^{**}$, two tailed t-test.

Word rareness We use a reference corpus consisting of ~ 400 million tweets to measure word frequencies. By computing a frequency ratio for each pair, we measure which word from a pair is more frequently used overall.

Perceived happiness To obtain happiness ratings for words we use the Hedonometer (Dodds et al. 2011; 2015). This consists of crowdsourced happiness ratings for 10,221 of the most frequent English words. The ratings range between 8.5 and 1.3 ($m = 5.37$, $\sigma = 1.08$). Note these do not only infer the emotional polarity of words (e.g., 'happiness' is more positive than 'terror'), but also how words are perceived by the reader individually without text context (e.g., 'mommy' is perceived happier than 'mom').

Concreteness Concreteness evaluates the degree to which the concept denoted by a word refers to a perceptible entity (Brysbaert, Warriner, and Kuperman 2013). Although the paraphrase pairs refer to the same entity, some words are perceived as more concrete (or conversely more abstract) than others. The dual-coding theory posits that concrete words are more easily learned, remembered and processed than abstract words (Paivio 1971). We use a list of 37,058 English words with crowdsourced ratings of concreteness on a scale from 5 (e.g., 'tiger' – 5) to 1 (e.g., 'spirituality' – 1.07) introduced in (Brysbaert, Warriner, and Kuperman 2013).

Results

The Pearson correlation results are shown in Table 9. We observe there are significant differences in paraphrase choice between user groups.

Males use words with slightly more characters and syllables, a fact previously claimed in (Mehl and Pennebaker 2003). On the other hand, females use more concrete options within paraphrase pairs. Previous studies (Newman et al. 2008) state that females are more likely to write about psychological states, while males are more likely to write about conceptual things, all of which are related to abstractness.

Most intriguingly, males choose the least happier word in a paraphrase pair. It is commonly stated (Mehl and Pennebaker 2003) that males both use less references to positive emotions and use more negative emotion words. Here, we show that this holds for words which are not necessarily a direct expression of an emotion. We also see that the mean pair happiness is correlated with the female score ($r = 0.068$, $p < .001$) and with the difference in happiness ($r = 0.027$, $p < .01$) i.e., pairs with higher happiness are both more polarizing between genders and, to a lesser degree, register larger differences in happiness. Figure 1 presents all paraphrase pairs with preserving pair order (first word is more feminine). We observe that most points have both very similar x-axis score (not polarizing enough between genders) and similar y-axis scores (it is expected that paraphrases which express the same concept to have similar happiness ratings). We highlight word pairs that are illustrative for all cases e.g., 'dip' is more feminine than 'decline' and is perceived happier (or not as bad in this case), while 'gorgeous' and 'super' are perceived similarly happy, but are polarize across genders. Most importantly, there is a significant correlation between female score and perceived happiness as illustrated by the positive slope of the correlation line.

For age, we observe that the older users prefer words that are longer in number of characters and syllables. This is expected, as vocabulary complexity increases with age. For happiness, we observe that younger users prefer words that sound happier, albeit less strongly correlated than in the case of gender. This is against previous psychology research that mentions an ageing positivity bias (Kern et al. 2014). Younger users also prefer words which are more concrete.

Regarding occupational class, we show that users with higher class use both longer words and words with more syllables, showing a higher language complexity. We note that these are the highest correlations for these two measures, demonstrating that occupational class influences word and syllable length more than gender or age. Users with a lower occupational class choose words which are happier and, to a much larger extent, more concrete. According to the model for psychological distance (Trope and Liberman 2010), abstract thinking is associated with how far people get from their own, immediate circumstances. As higher occupational class users use Twitter more professionally and less interpersonally (Preoţiuc-Pietro et al. 2015), the word choice of these users is more abstract.

Rare word usage is associated with females and, to a lesser extent, with younger and lower skilled users. Rare word usage is likely to conflate lexical innovation, non-standard forms and specialised in-vocabulary words. To investigate in more detail, we measure diversity in language choice. We compute the average entropy of word use within a cluster for each attribute, with higher entropy if the phrase choice distribution is more uniform and thus, more diverse. Using the entropy, we uncover that higher skilled users are most diverse ($\mu = .794$), followed by males ($\mu = 0.779$) and older users ($\mu = 0.765$). Females ($\mu = 0.753$), lower skilled users ($\mu = 0.742$) and, most importantly, younger users ($\mu = 0.683$) have lower entropies, revealing that they are less flexible in phrase choice.

Simulating Writing Style

We consider a interesting application of our work to be forging gender or other traits by manipulating phrase choice. As a proof of concept, we build upon previous work of stylistic paraphrasing (Xu et al. 2012; Mizukami et al. 2015), which

Female Male	@USER A teeny tiny cute coffee shop (and a cheap fry-up style cafe). I'm intrigued... @USER A small tiny clever coffee shop (and a cheap fry-up style cafe). I'm intrigued...
Female Male	Well, this gorgeous video of St. Petersburg got me very excited or my trip (2 weeks to go!) Well, this superb video of St. Petersburg got me prohibitively aroused or my journey (2 weeks to go!)
Older Younger	hilton worldwide launches its largest global career event URL #csr hilton worldwide starts its biggest global career event URL #csr
Older Younger	rt @USER : how your body responds to exercise over time URL rt @USER : how your body answers to workout over time URL

Table 10: Example of tweets and their automatically generated counterparts. Can you guess which are the original tweets?

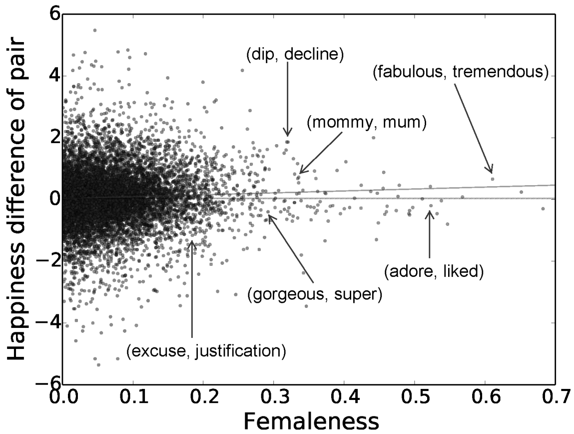

Figure 1: Scatter plot showing best fit line between gender paraphrase use differences and differences in perceived word happiness. Each dot is a paraphrase pair. The first word in pair is the more feminine one.

was originally conceived for translation from contemporary modern English to Shakespearean style, or for translation of spoken text towards an individual speaker's style. We use the phrase-based statistical machine translation decoder Moses (Koehn et al. 2007) to combine a translation model of the stylistic paraphrases uncovered by our scoring function and a language model of the targeted demographic group. This method essentially performs lexical substitution while taking context into account. Table 10 shows examples that substitute the words highly scored of one gender for counterparts of the opposite gender. The original tweets are the top ones and their automatically generated counterparts the bottom ones; it's not easy to tell apart the 'fake' ones and most preserve grammaticality and meaning.

Although this preliminary system can successfully alter the style of tweets and preserve grammaticality within certain bounds, we found that perceptions are still dominated by topical information and are hard to rate individually for user traits. We also found writing style of a demographic group (gender, occupational class and age) to be more blurry and subtle than that of a specific speaker (Mizukami et al. 2015) or a specific writer (Xu et al. 2012) than were captured by language models in previous work. A more sophisticated text-to-text generation system use parallel data (Xu, Ritter, and Grishman 2013) and feature-rich paraphrase rules may help solve these two problems (Ganitkevitch et al. 2011) and thus leave this for future work.

Conclusions

User attribute differences exist at a stylistic dimension. We have presented the first attempt to capture these by using a method build on paraphrase choice combined with statistics computed from vast volumes of social media text. We have shown that significant changes exist at the phrase choice level and these are both predictive of the user traits and intuitive to human annotators. Our analysis can be generalised easily to other user attributes for which large user-centered data sets are available such as location (Cheng, Caverlee, and Lee 2010) or personality (Schwartz et al. 2013).

Our methods can be further improved by using better paraphrasing (Wieting et al. 2015) or word senses (Apidianaki, Verzeni, and McCarthy 2014) to adjust for paraphrases which may be skewed by a sense which is topical e.g., 'collapse' – 'crumble'. We have used PPDB 2.0, a general purpose paraphrase resource for computing paraphrase pairs. High-quality large-scale Twitter-specific paraphrase resources are still scarce, but methods such as (Xu, Ritter, and Grishman 2013) should provide a more detailed view of stylistic difference. A future model can also include attribute labels as a variable in the paraphrase extraction phase. In this study we have solely relied on statistics computed over very large data sets.

Acknowledgments

Daniel Preoţiuc-Pietro and Lyle Ungar acknowledge the support of the Templeton Religion Trust, grant TRT-0048. Wei Xu acknowledges the support of the NSF grant IIS-1430651 and the Google Faculty Research Award awarded to Chris Callison-Burch. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of the NSF or the U.S. Government. We thank the reviewers and the senior program chair for valuable feedback.

References

Apidianaki, M.; Verzeni, E.; and McCarthy, D. 2014. Semantic clustering of pivot paraphrases. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC, 4270–4275.

Bamman, D.; Eisenstein, J.; and Schnoebelen, T. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2):135–160.

Barzilay, R. 2003. *Information fusion for multidocument summarization: Paraphrasing and generation*. Ph.D. Dissertation.

Bergsma, S., and Van Durme, B. 2013. Using conceptual class attributes to characterize social media users. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL, 710–720.

Brysbaert, M., and Warriner, A.B. and Kuperman, V. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* , 46(3):904–911.

Burger, D. J.; Henderson, J.; Kim, G.; and Zarrella, G. 2011. Discriminating gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP, 11–1309.

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You are where you tweet: A content-based approach to geo-locating Twitter users. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management*, CIKM, 759–768.

Danescu-Niculescu-Mizil, C.; Lee, L.; Pang, B.; and Kleinberg, J. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, WWW, 699–708.

Derczynski, L.; Ritter, A.; Clark, S.; and Bontcheva, K. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP, 198–206.

Dodds, P. S.; Harris, K. D.; Kloumann, I. M.; Bliss, C. A.; and Danforth, C. M. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS One* 6 (12).

Dodds, P. S.; Clark, E. M.; Desu, S.; Frank, M. R.; Reagan, A. J.; Williams, J. R.; Mitchell, L.; Harris, K. D.; Kloumann, I. M.; Bagrow, J. P.; Megerdoomian, K.; McMahon, M. T.; Tivnan, B. F.; and Danforth, C. M. 2015. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences* 112(8):2389–2394.

Eisenstein, J.; O'Connor, B.; Smith, N. A.; and Xing, E. P. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP, 1277–1287.

Ganitkevitch, J.; Callison-Burch, C.; Napoles, C.; and Van Durme, B. 2011. Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP, 1168–1179.

Ganitkevitch, J.; Van Durme, B.; and Chris Callison-Burch, C. 2013. PPDB: The paraphrase database. In *Proceedings of the 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL, 758–764.

Mingkun Gao and Wei Xu and Chris Callison-Burch 2015. Cost Optimization in Crowdsourcing Translation. In *Proceedings of the 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL, 705–713.

Haas, A. 1979. Male and female spoken language differences: Stereotypes and evidence. *Psychological Bulletin* 86:616–626.

Han, B.; Cook, P.; and Baldwin, T. 2012. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP, 421–432.

Hovy, D. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, ACL, 752-762.

Kern, M. L.; Eichstaedt, J. C.; Schwartz, H. A.; Park, G.; Ungar, L. H.; Stillwell, D. J.; Kosinski, M.; Dziurzynski, L.; and Seligman, M. E. 2014. *Developmental Psychology* 50:178–188.

Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the Association for Computational Linguistics (Demo)*, 177–180.

Lampos, V.; Aletras, N.; Preoţiuc-Pietro, D.; and Cohn, T. 2014. Predicting and characterising user impact on Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, 405–413.

Lui, M., and Baldwin, T. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ACL, 25–30.

Mehl, M. R., and Pennebaker, J. W. 2003. The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology* 84(4):857–870.

Mizukami, M.; Neubig, G.; Sakti, S.; Toda, T.; and Nakamura, S. 2015. Linguistic individuality transformation for spoken language. In *Natural Language Dialog Systems and Intelligent Assistants*, 129–143.

Newman, M. L.; Groom, C. J.; Handelman, L. D.; and Pennebaker, J. W. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes* 45(3):211–236.

Paivio, A.; 1971. Imagery and verbal processes. *New York: Holt, Rinchart, and Winston*.

Pavlick, E., and Nenkova, A. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *Proceedings of the 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL, 218–224.

Pavlick, E.; Bos, J.; Nissim, M.; Beller, C.; Van Durme, B.; and Callison-Burch, C. 2015. Adding semantics to data-driven paraphrasing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, ACL, 1512–1522.

Preoţiuc-Pietro, D.; Samangooei, S.; Cohn, T.; Gibbins, N.; and Niranjan, M. 2012. Trendminer: an architecture for real time analysis of social media text. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media, Workshop on Real-Time Analysis and Mining of Social Streams*, ICWSM, 38–42.

Preoţiuc-Pietro, D.; Volkova, S.; Lampos, V.; Bachrach, Y.; and Aletras, N. 2015. Studying user income through language, behaviour and affect in social media. *PLoS ONE* 10 (9).

Preoţiuc-Pietro, D.; Lampos, V.; and Aletras, N. 2015. An analysis of the user occupational class through Twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, ACL.

Rao, D.; Yarowsky, D.; Shreevats, A.; and Gupta, M. 2010. Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, SMUC, 37–44.

Sap, M.; Park, G.; Eichstaedt, J.; Kern, M.; Ungar, L.; and Schwartz, H. A. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP, 1146–1151.

Sarawgi, R.; Gajulapalli, K.; and Choi, Y. 2011. Gender attribution: Tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNNL, 78–86.

Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E.; and Ungar, L. H. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS One*, 8 (9).

Trope, Y., and Liberman, N. 2010. Construal-level theory of psychological distance. *Psychological Review*, 117(2), 440.

Van Durme, B. 2012. Streaming analysis of discourse participants. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP, 48–58.

Vogel, A., and Jurafsky, D. 2012. He said, she said: Gender in the acl anthology. In *Proceedings of the Special Workshop on Rediscovering 50 Years of Discoveries*, ACL, 33–41.

Volkova, S., and Bachrach, Y. 2015. On predicting sociodemographic traits and emotions in social networks and implications to online self-disclosure. *Cyberpsychology, Behavior, and Social Networking*, To appear.

Volkova, S.; Coppersmith, G.; and Van Durme, B. 2014. Inferring user political preferences from streaming communications. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL, 186–196.

Volkova, S.; Wilson, T.; and Yarowsky, D. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP, 1815–1827.

Wieting, J.; Bansal, M.; Gimpel, K.; and Livescu, K. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, TACL, 345–358.

Xu, W.; Ritter, A.; Dolan, B.; Grishman, R.; and Cherry, C. 2012. Paraphrasing for style. In *Proceedings of the 24th International Conference on Computational Linguistics*, COLING, 2899–2914.

Xu, W.; Ritter, A.; and Grishman, R. 2013. Gathering and generating paraphrases from Twitter with application to normalization. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, BUCC, 121–128.