Personality Driven Differences in Paraphrase Preference

Daniel Preoţiuc-Pietro
Positive Psychology Center
University of Pennsylvania
danielpr@sas.upenn.edu

Jordan Carpenter
Kenan Institute for Ethics
Duke University
jmc51@duke.edu

Lyle Ungar
Computer and Information Science
University of Pennsylvania
ungar@cis.upenn.edu

Abstract

Personality plays a decisive role in how people behave in different scenarios, including online social media. Researchers have used such data to study how personality can be predicted from language use. In this paper, we study phrase choice as a particular stylistic linguistic difference, as opposed to the mostly topical differences identified previously. Building on previous work on demographic preferences, we quantify differences in paraphrase choice from a massive Facebook data set with posts from over 115,000 users. We quantify the predictive power of phrase choice in user profiling and use phrase choice to study psycholinguistic hypotheses. This work is relevant to future applications that aim to personalize text generation to specific personality types.

1 Introduction

The task of user trait prediction from text has increased in popularity and importance with the availability of user generated content which encodes various information about the author of the text. Using machine learning techniques and large data sets, past research managed to predict with varying degrees of accuracy a series of both demographic traits such as age (Rao et al., 2010; Sap et al., 2014), gender (Burger et al., 2011; Rangel et al., 2015; Flekova et al., 2016a), location (Eisenstein et al., 2010), political affiliation (Volkova et al., 2014; Preoţiuc-Pietro et al., 2017), popularity (Lampos et al., 2014), occupation (Preoţiuc-Pietro et al., 2015b; Liu et al., 2016), income (Preoţiuc-Pietro et al., 2015c; Flekova et al., 2016b) and psychological traits such as personality dimensions (Schwartz et al., 2013; Preoţiuc-Pietro et al., 2016a) or mental states (De Choudhury et al., 2013; Coppersmith et al., 2014; Preoţiuc-Pietro et al., 2015a).

For psychological traits of users, a key set of traits is represented by personality, with the Five Factor Model or the 'Big Five' being the most widely used model for representing personality. This posits the existence of five traits in which people vary: openness to experience, conscientiousness, extraversion, agreeableness and neuroticism (McCrae and John, 1992). Methods for user trait prediction can uncover sociological insight into user behaviour or implicit biases and also improve a range of applications in recommender systems, targeted marketing or in natural language processing where they can lead to improvements in tasks such as text classification (Hovy, 2015) or sentiment analysis (Volkova et al., 2013). While these methods achieve good predictive performance, they pose significant challenges to the anonymization of identity online.

Most differences in language use across traits are topical. For example, users high in extraversion post more about social activities ('party', 'cant wait', 'weekend'), while introverts prefer to post more about computer related activities ('Internet', 'computer', 'anime'). Users high in neuroticism post about their negative feelings ('depressed', 'sick of', 'lonely'), while users low in neuroticism post more about religion ('blessings', 'praise') or sports ('basketball', 'soccer', 'success') (Park et al., 2015).

However, stylistic rather than topical differences are needed in some applications. For example, (Mirkin et al., 2015) propose that the output text of machine translation systems should reproduce the traits of the author of the source text. In this case, topical information is fixed, and the trait information can be transmitted only using stylistic cues. Following the work of (Preoţiuc-Pietro et al., 2016b) who studied demographic traits, we

Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science, pages 17–26,
Vancouver, Canada, August 3, 2017.

study in this paper user personality differences in paraphrase choice – a specific type of stylistic difference. Paraphrases represent alternative ways to convey the same information (Barzilay, 2003), using either single words or short phrases. Table 1 presents a couple of motivating examples of two group of words and phrases which are all paraphrases of each other ordered by the frequency of use for each personality trait.

In this study, we measure for the first time the differences in paraphrase usage between personality types from a large social media data set in an attempt to obtain language differences isolated from topical influence. Our analysis measures similarities between personality traits, the predictive power of stylistic words and a number of psycholinguistic theories about word choice. The paraphrase scores for each of the five personality traits are available online.[1]

2 Data

Our complete data set consists of approximately 15 million Facebook status updates posted by 115,312 users, representing the full MyPersonality data set (Kosinski et al., 2013). Participants volunteered to share their status updates as part of the MyPersonality application, providing informed consent for data collection. In the MyPersonality application they took a variety of questionnaires, including the International Personality Item Pool proxy for the NEO Personality Inventory Revised (NEO-PI-R) (McCrae and John, 1992; Costa and McCrae, 2008), based on which the five personality trait scores are computed for each user (ranging from 1 to 5).

We split our users into binary groups for each personality trait. In order to have non-overlapping groups, we selected the top 20% users as being high in one trait and the bottom 20% as low in that trait. Data set statistics are presented in Table 2. Our methodology requires a split of users into dichotomous groups in order to compute paraphrase preference. We acknowledge that this split represents a simplification of personality traits and of the subsequent personality prediction task, although this was also used in some previous research (Mairesse et al., 2007; Celli et al., 2014) and, due to the ordinal nature of the personality scores, is highly unlikely to qualitatively affect our results.

[1] http://www.preotiuc.ro

Personality Trait	Low	High
Openness	≤ 3.25 (25,211 users)	≥ 4.5 (24,700 users)
Conscientiousness	≤ 2.75 (23,221 users)	≥ 4.049 (23,639 users)
Extroversion	≤ 2.75 (23,802 users)	≥ 4.25 (26,310 users)
Agreeableness	≤ 3 (27,723 users)	≥ 4.25 (23,750 users)
Neuroticism	≤ 2 (25,798 users)	≥ 3.5 (23,339 users)

Table 2: Personality score thresholds and number of users in each personality trait group for the analysis.

3 Quantifying Personality Differences

We use the Paraphrase Database (PPDB) (Ganitkevitch et al., 2013) as our source of paraphrases, owing to its very large size and quality. PPDB 2.0 (Pavlick et al., 2015b) contains 23.820.422 paraphrases derived from a large collection of bilingual texts by pivoting methods. The phrases part of paraphrases are up to three tokens in length (1–3 grams). In PPDB 2.0, each paraphrase pair comes with predicted scores for the relation type between the two phrases ('Equivalence', 'Entailment', 'Exclusion', 'Other relation', 'Unrelated') obtained using a supervised regression model using lexical, distributional and other features (Pavlick et al., 2015a). While there is no inarguable definition of the paraphrase term (Androutsopoulos and Malakasiotis, 2010; Bhagat and Hovy, 2013), in this work we are most interested in the most restrictive type of relationship ('Equivalence') as described in (Pavlick et al., 2015a). We thus use paraphrase pairs that have an equivalence score of at least 0.2 (chosen based upon the inspection of the pairs), leaving us with 6.157.570 paraphrase pairs.

Given a paraphrase pair, we use phrase occurrence statistics computed over our data set to measure the phrase choice difference over user attributes. For the rest of this paragraph, we exemplify with the trait of extraversion, but the computation is analogous for the other four traits.

To score how much a user group favors a phrase w, we compute the scores Extravert(w) and Introvert(w). These are computed by counting the number of times phrase w was used by a user divided by the total number of words of that used, then averaging across all users high or low extraversion respectively. For each phrase we then compute a score:

$$\text{Extraversion}(w) = \log\left(\frac{\text{Extravert}(w)}{\text{Introvert}(w)}\right) \quad (1)$$

Within a paraphrase pair (w_1, w_2), the difference Extraversion(w_1)−Extraversion(w_2) measures the

Low	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
↓	firstly (-1.24)	firstly (-0.62)	above all (-0.30)	first of all (-0.20)	foremost (-0.24)
	first (-1.03)	foremost (-0.23)	firstly (-0.13)	first (-0.09)	most importantly (-0.21)
	foremost (-0.47)	first of all (-0.20)	first (-0.11)	foremost (-0.07)	above all (-0.08)
	first of all (0.49)	first (0.00)	foremost (-0.07)	above all (0.03)	first (0.01)
	most of all (0.59)	above all (0.14)	most of all (0.10)	firstly (0.14)	most of all (0.02)
	most importantly (0.79)	most importantly (0.16)	first of all (0.26)	most importantly (0.19)	first of all (0.03)
	above all (0.86)	most of all (0.42)	most importantly (0.48)	most of all (0.21)	firstly (0.40)
High	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**

Low	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
↓	Stunning (-.45)	Magnificent (-1.12)	Excellent (-.42)	Marvelous (-.85)	Tremendous (-.57)
	Great (-.34)	Awesome (-.40)	Splendid (-.37)	Unbelievable (-.51)	Remarkable (-.28)
	Wonderful (-.20)	Super (-.36)	Marvelous (-.31)	Remarkable (-.25)	Terrific (-.22)
	Magnificent (-.18)	Splendid (-.30)	Awesome (-.26)	Stunning (-.17)	Marvelous (-.17)
	Super (-.18)	Amazing (-.21)	Exciting (-.14)	Excellent (-.16)	Unbelievable (-.09)
	Gorgeous (-.12)	Excellent (-.12)	Fantastic (-.10)	Super (-.10)	Incredible (-.08)
	Exciting (-.10)	Stunning (-.08)	Great (-.07)	Gorgeous (-.09)	Fabulous (-.07)
	Fabulous (-.09)	Gorgeous (-.08)	Wonderful (-.07)	Awesome (.00)	Awesome (-.03)
	Amazing (-.07)	Incredible (-.04)	Super (-.04)	Fabulous (.02)	Excellent (-.03)
	Tremendous (-.04)	Exciting (.00)	Incredible (-.02)	Amazing (.05)	Great (-.02)
	Awesome (-.02)	Unbelievable (.03)	Unbelievable (-.02)	Great (.07)	Wonderful (-.02)
	Unbelievable (.00)	Fantastic (.07)	Remarkable (-.01)	Fantastic (.10)	Exciting (.01)
	Fantastic (.03)	Great (.18)	Amazing (.07)	Incredible (.18)	Fantastic (.02)
	Marvelous (.13)	Fabulous (.23)	Terrific (.07)	Exciting (.19)	Splendid (.05)
	Terrific (.22)	Wonderful (.38)	Gorgeous (.12)	Terrific (.27)	Super (.06)
	Incredible (.22)	Terrific (.39)	Stunning (.19)	Tremendous (.31)	Amazing (.09)
	Splendid (.29)	Marvelous (.50)	Magnificent (.31)	Wonderful (.35)	Gorgeous (.29)
	Excellent (.36)	Remarkable (.55)	Fabulous (.37)	Splendid (.39)	Magnificent (.44)
	Remarkable (.61)	Tremendous (.70)	Tremendous(.72)	Magnificent (.51)	Stunning (.57)
High	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**

Table 1: Two example groups of phrases that are all paraphrases of each other. Words and phrases are ordered by frequency of use. The top words are more frequently used by users low in each personality trait, with words further down the list being more specific of users high in the respective personality trait. The number in brackets represents the score with which the word is related to each trait (described in Section 3).

stylistic distance between users high in extraversion compared to users low in extraversion. This method of computing stylistic distance is similar to the work of Pavlick and Nenkova (2015) who studied paraphrasing in the context of formality and complexity and to that of Preoţiuc-Pietro et al. (2016b) who looked at differences between gender, age and occupational class groups.

In a few experiments, we also use paraphrase clusters which are created by using the transitive closure of pairwise paraphrases, as the supervised model for scoring equivalence combined with our threshold leads to transitivity not holding in our list of pairs. Within these clusters, we subtract the mean phrase score to adjusts for topic prevalence and to lead to a score of 0 representing a point of alignment across all clusters. In total, we derive 785.226 paraphrase clusters (mean = 7.43 words, median = 4 words, st.dev = 11.06 words). Out of these, on average 171.788 clusters (mean = 5.20 words) across the five personality traits contain at least two words scored for phrase choice, as we remove words with low frequency in our data (a relative frequency of under 10^{-5} in our data set).

4 Predicting Personality

We first test the predictive power of paraphrases in the prediction task of whether a user is high or low in each personality trait. We randomly select 90% of the users to build the scores for all phrases and keep 10% of users for evaluating prediction accuracy. We use the Naïve Bayes classifier to assign a score to each user. We use this classifier as this computes for each word the log probability of the word belonging to one class (similar to the measure we previously defined) and computes the

dot product between this distribution and the user phrase frequency vector. We chose this algorithm over others to directly tests the viability of our metric. The prior class distribution is estimated based on the training data and we use Laplace smoothing.

To measure the influence of paraphrase choice, we compare the performance of the model using only phrases appearing in at least one paraphrase pair (a proxy for stylistic choice, 62.919 phrases), the rest of the phrases separately (a proxy for topical information, 54.197 phrses) as well as the combined set of phrases. The vocabulary consists of 117.117 phrases (1–3 grams) which have a relative frequency of over 10^{-5} in our data set. Results on predicting personality for unseen users measured in accuracy are shown in Table 3.

	Ope	Con	Ext	Agr	Neu
Random Baseline	.500	.500	.500	.500	.500
Only Paraphrases	.603	.551	.519	.551	.549
Phrases w/o Paraphrases	.573	.589	.578	.553	.590
All Phrases	**.623**	**.639**	**.597**	**.593**	**.631**

Table 3: User attribute prediction results evaluated in accuracy. Using only paraphrases that capture more stylistic rather than topical differences between different personality trait groups, our method still shows good predictive power comparing to using all phrase (1–3 grams) features.

We notice that overall personality can be predicted with significant margins even when using a simple Naive Bayes approach without any feature selection. Both phrases part of paraphrase pairs and not part of paraphrase pairs significantly improve on the random baseline with one exception (Extraversion and paraphrases). However, the numbers are lower than in the case of user demographics (Preoţiuc-Pietro et al., 2016b), which is to be expected when predicting psychological traits (Schwartz et al., 2013; Rangel et al., 2015).

We highlight that in the case of openness to experience, the phrases that are part of paraphrase pairs obtain better prediction performance in accuracy than the other set of phrases. The latter perform better when predicting conscientiousness, extraversion and neuroticism and comparable in case of agreeableness. Combining all phrases consistently obtains the best results.

5 Trait Differences

A very revealing aspect of paraphrase choice for each trait is the order of preference within a paraphrase cluster, as exemplified in Table 1. To quantify this preference across all clusters, we compute the cluster rank similarity between all pairs of user traits. The average Kendall τ rank correlation coefficient across all clusters is presented in Table 4. As certain personality trait scores are correlated and some users might be part of multiple groups, we also show the correlations between the trait scores in Table 5. As the number of users is very large ($>$100.000), all correlations in Tables 4 and 5 are significant.

The results on paraphrase choice show a few distinctive patterns. In both paraphrase choice and actual personality scores, neuroticism is anti-correlated with all other four traits, albeit more strongly in case of personality scores. Openness to experience is weakly negatively correlated with all four traits in paraphrase choice, while it is overall weakly positively correlated with the other traits in personality scores. Paraphrase choice is positively correlated across the other three traits (conscientiousness, extraversion, agreeableness), similarly to actual personality scores and with comparable correlations numbers.

Overall, this analysis demonstrates that overall, stylistic paraphrase choice largely reflects user level differences with some variation in case of openness to experience.

	Ope	Con	Ext	Agr	Neu
Ope	–	-.071	-.018	-.040	-.028
Con	-.071	–	.134	.174	-.211
Ext	-.028	.134	–	.107	-.180
Agr	-.040	.174	.107	–	-.174
Neu	-.028	-.211	-.180	-.174	–

Table 4: Average Kendall τ rank correlation between paraphrase cluster usage compared across different user traits. Spearman rank correlation and Pearson correlation reveal similar patterns.

	Ope	Con	Ext	Agr	Neu
Ope	–	.031	.129	.039	-.047
Con	.031	–	.192	.177	-.303
Ext	.129	.192	–	.169	-.337
Agr	.039	.177	.169	–	-.326
Neu	-.047	-.303	-.337	-.326	–

Table 5: Correlation between personality traits in our data set.

6 Linguistic Hypotheses

We investigate a number of psycholinguistic hypotheses about language choice and style by using our paraphrase based method. We argue that word choice within a paraphrase pair excludes the topical influence that confounds studies using all words (Sarawgi et al., 2011)

6.1 Word Properties

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 simple proxy for word complexity.

Word Rareness To measure word frequency, we use a reference corpus retrieved from the 10% sample of the Twitter stream between 2 January – 28 February 2011 ($\sim$ 400 million tweets), filtered for English using the Trendminer pipeline (Preoţiuc-Pietro et al., 2012). We measure which word from a pair is more frequently used overall by computing a ratio between the frequencies of the two words within a pair.

Perceived Happiness We use the Hedonometer (Dodds et al., 2011, 2015) to obtain happiness ratings for single words. The Hedonometer consists of crowdsourced happiness ratings for 10,221 of the most frequent English words. The ratings range between 8.5 and 1.3 ($\mu = 5.37$, $\sigma = 1.08$). 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'). We compare the user group preference with the difference in happiness ratings.

Affective Norms To compliment the happiness ratings, we use information about the affective norms of words. In the dimensional model of emotions, any particular emotion can be defined as a set of values on a number of different dimensions. One of the most popular models consists of three dimensions (Mehrabian and Russell, 1974): **Valence** – pleasant vs. unpleasant; **Arousal** – excited vs. calm; **Dominance** – controlled vs. in-control. We use a list of $\sim$14,000 words rated in all three affective norms introduced in (Warriner et al., 2013). For words rated in both perceived happiness and valence, the correlation is very high ($r = .918$).

Concreteness Concreteness evaluates the degree to which the concept denoted by a word refers to a perceptible entity (Brysbaert et al., 2014). 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 humans process and represent verbal and non-verbal information in separate, related systems. According to this, both concrete and abstract words are represented in the verbal system, but only concrete words are represented in the non-verbal system. Thus, concrete words are more easily learned, remembered and processed than abstract words (Paivio, 2013). We use a list of 37,058 English words with ratings of concreteness on a scale from 5 (e.g., 'tiger' – 5) to 1 (e.g., 'spirituality' – 1.07) introduced in (Brysbaert et al., 2014).

Imageability The construct of imageability represents how easily a particular word elicits a mental picture of the word's referent (Toglia and Battig, 1978). Imagery is thought to be an important aspect of the non-verbal system in the dual-coding theory and is correlated with concreteness ($r = .78$) (Gilhooly and Logie, 1980). We use 6,000 ratings on the ease or difficulty with which words arouse mental images for mono- and disyllabic words (Cortese and Fugett, 2004; Schock et al., 2012), ranging from e.g., 1.2 – 'an' to 7 – 'blizzard'.

Sensory Experience Sensory experience ratings reflect the extent to which a word evokes a sensory and/or perceptual experience in the mind of the reader (Juhasz and Yap, 2013). In contrast to imageability which explicitly refers to visual and sound images and asks raters to attempt to build a mental image of the concept, the sensory experience ratings measures the ability for a word to evoke an actual sensation (taste, touch, sight, sound, or smell) that occurs when reading the word. Although sensory experience and imageability are correlated ($r = .586$) (Juhasz and Yap, 2013), the two variables independently predict unique variance in lexical-decision latencies (Juhasz et al., 2011). We use the ratings from (Juhasz and Yap, 2013) which consist of 5,000 word ratings (e.g., 1 – 'those'; 3 –

Feature	Ope	Con	Ext	Agr	Neu
Word length	$.182^{**}$	$.097^{**}$	$.080^{**}$	.010	$-.065^{**}$
#Syllables	$.067^{**}$	$.045^{**}$	$.047^{**}$	$.016^{*}$	$-.020^{*}$
Word rareness	$-.022^{**}$	.005	$.013^{*}$	.007	$-.004$
Happiness	$-.027^{*}$	$.039^{**}$	$.034^{*}$	$.040^{**}$	.004
Valence	$-.041^{**}$	$.050^{**}$	$.050^{**}$	$.054^{**}$	.006
Arousal	$-.012$	$-.001$	$.028^{*}$	.005	$-.024^{*}$
Dominance	$-.043^{**}$	$.036^{**}$	$.031^{*}$	$.030^{*}$	.000
Concreteness	$-.068^{**}$	$-.014$	.010	$-.007$	$.023^{*}$
Imageability	$-.061^{*}$	$-.010$	.026	.027	.016
Sensory Experience	$-.010$	$-.018$	.023	.001	$.064^{**}$
Age-of-Acquisition	$.163^{**}$	$-.002$	$-.060^{**}$	$-.032^{*}$	$-.014$

Table 6: Correlation coefficients between word property differences and word preference by users high in each personality trait across all paraphrase pairs – $p < 0.05$, two tailed t-test, significant after false discovery rate multi-comparison corrections: Benjamini-Hochberg (*), Bonferroni (**).

'relief'; 6 – 'music').

Age-of-Acquisition Age-of-Acquisition is a psycholinguistic variable referring to the age at which a word is typically learned (Kuperman et al., 2012). Words with higher age-of-acquisition are anti-correlated to sensory experience ($r = -.586$), imageability ($r = -.440$) (Juhasz and Yap, 2013) and correlated with length in letters ($r = .549$), syllables ($r = .528$) and, to a lesser extent, to abstractness ($r = .166$) (Kuperman et al., 2012). We use the age-of-acquisition ratings for 30,000 words rated with the year in which the words are acquired (e.g., 'momma' – 1.58; 'foot' – 3.44; 'bipartisan' – 16.2) introduced in (Kuperman et al., 2012).

6.2 Paraphrase Entropy

Additionally, we are interesting in identifying which personality groups prefer using a more diverse set of alternative phrases, rather than using a few idiosyncratic phrases. Using all paraphrase clusters (1–3 grams), we compute the average entropy over paraphrase cluster distributions. A higher entropy means the distribution is less peaked towards a specific word, thus showing higher variety in choice.

6.3 Results

We establish if a group of users prefers words within paraphrase pairs with one of the characteristics presented in the previous section using the following method. For each trait and paraphrase pair, we compute the stylistic difference between the words within a pair (see Section 3). Then, for each trait, we run a Pearson correlation between the vector of stylistic difference scores for each pair and the vector containing the differences in word characteristics (e.g. the difference between the number of syllables of the two words). For each word property, we only retain the paraphrase pairs where we can measure both words, which leads to different numbers of pairs (and hence difference significance thresholds) for each test. The Pearson correlation results are shown in Table 6. We observe there are several statistically significant differences in paraphrase choice between the user groups. Paraphrase entropy by personality trait groups are presented in Table 7.

Personality Trait	Low	High
Openness $^{(**)}$	.838	.924
Conscientiousness	.893	.894
Extroversion $^{(**)}$	.901	.891
Agreeableness $^{(*)}$	.899	.894
Neuroticism $^{(**)}$	.900	.892

Table 7: Average paraphrase cluster entropies for each personality trait. The higher the entropy, the more diverse is the paraphrase choice of the specific group of users. Mean differences are tested for significance using the Mann-Whitney Test: $p \leq .05^{(*)}$, $p \leq .001^{(**)}$.

The trait that leads to the largest number of significant correlations with phrase choice is openness to experience. Users high in openness prefer words which are longer and with more syllables. These patterns are consistent with the theory that open people are intellectually attuned, creative, and curious (McCrae and Costa Jr, 1997). Simultaneously,

openness to experience was negatively related to concreteness, dominance, valence and happiness. This indicates that users who are high in openness are more likely to express themselves in indirect and abstract ways, and they are less likely to prefer explicitly happier words. Again, these are consistent with a more cerebral or artistic mode of communication. Word rareness is anti-correlated with high in openness. However, we noticed that word rareness captures in a large extent also misspellings and alternative spellings. In terms of entropy however, openness to experience generates by far the largest difference in group means for entropy. Those interested in novelty and new experiences may especially dislike phrasing the same concept in the same way over time when other options are available, prefer idiosyncratic words and may have larger vocabularies.

Conscientiousness, extraversion and agreeableness have similar correlations across all phrase choice traits. Users high in these three traits prefer words that are longer and have more syllables. However, for extraversion and agreeableness, age-of-acquisition results show that these groups tend not to choose words acquired later and entropy results show a more limited breadth in usage, both indicative of less complex word choice. Especially, introverts score higher in these choices, perhaps because introverts prefer solitary activities such as reading and may therefore have larger and more sophisticated vocabularies (Furnham, 1981).

All three traits prefer happier and more dominant words, which, at least for extraversion, is unsurprising as these qualities are part of the definition of the trait (Watson and Clark, 1997). Users high in agreeableness are also known to express higher positive valence and conscientious users tend to be more dominant.

Despite the opposite patterns in language use associated with these three traits and openness, these are positively correlated in the user population. Therefore, the two sets of correlations are not simply the same effect explained in two different ways.

Neuroticism exhibits the fewest correlations with phrase choice. Users high in this trait prefer words that are shorter, have fewer syllables and have a slightly lower entropy, which indicates a mild tendency for simpler, idionsyncratic words. Finally, users high the neuroticism prefer words that are higher in sensory experience, and to a lesser degree, that are more concrete. This underlines the preference of this group of users to use social media as a means of communicating about the immediate context.

7 Conclusions

We have studied phrase choice, a particular type of stylistic language difference, across the Big Five personality traits for the first time. We used a large data-driven paraphrase dictionary as our source of paraphrases in combination with statistics computed over large volumes of Facebook status updates. We have shown paraphrase words are, with one exception, predictive of the personality traits and that differences exist in phrase choices. Our analysis of several psycholinguistic word characteristics showed that personality correlates with many systematic word choices and these are intuitive and correspond to theories of personality.

Differences in paraphrase choice are likely to be useful in text-to-text generation and dialogues systems. Tailoring automatically generated text based on personality traits might be desirable in multiple scenarios, such as for tutoring or customer support. However, in most of these cases, the topic is fixed and personalization can be achieved only at a stylistic level. To this end, we make our scored paraphrase choices across personality traits publicly available.

Acknowledgments

The authors acknowledge the support of the Templeton Religion Trust, grant TRT-0048.

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A Survey of Paraphrasing and Textual Entailment Methods. *Journal of Artificial Intelligence Research* 38:135–187.

Regina Barzilay. 2003. *Information Fusion for Multidocument Summarization: Paraphrasing and Generation*. Ph.D. thesis, Columbia University.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics* 39(3):463–472.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness Ratings for 40 Thousand Generally known English Word Lemmas. *Behavior Research Methods* 46(3):904–911.

D. John Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on

Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. EMNLP, pages 1301–1309.

Fabio Celli, Elia Bruni, and Bruno Lepri. 2014. Automatic Personality and Interaction Style Recognition from Facebook Profile Pictures. In *Proceedings of the 22nd ACM International Conference on Multimedia*. MM, pages 1101–1104.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. In *Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (CLPsych)*. ACL, pages 51–60.

Michael J Cortese and April Fugett. 2004. Imageability Ratings for 3,000 Monosyllabic Words. *Behavior Research Methods, Instruments, & Computers* 36(3):384–387.

Paul T Costa and Robert R McCrae. 2008. The Revised NEO Personality Inventory (NEO-PI-R). *The SAGE Handbook of Personality Theory and Assessment* 2:179–198.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. ICWSM, pages 128–137.

Peter Sheridan Dodds, Eric M. Clark, Suma Desu, Morgan R. Frank, Andrew J. Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M. Kloumann, James P. Bagrow, Karine Megerdoomian, Matthew T. McMahon, Brian F. Tivnan, and Christopher M. Danforth. 2015. Human Language Reveals a Universal Positivity Bias. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 112(8):2389–2394.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PloS ONE* 6(12):e26752.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP, pages 1277–1287.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016a. Analyzing Biases in Human Perception of User Age and Gender from Text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. ACL, pages 843–854.

Lucie Flekova, Lyle Ungar, and Daniel Preoctiuc-Pietro. 2016b. Exploring Stylistic Variation with Age and Income on Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. ACL, pages 313–319.

Adrian Furnham. 1981. Personality and Activity Preference. *British Journal of Social Psychology* 20(1):57–68.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of the 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. NAACL, pages 758–764.

Ken J Gilhooly and Robert H Logie. 1980. Age-of-acquisition, Imagery, Concreteness, Familiarity, and Ambiguity Measures for 1,944 Words. *Behavior Research Methods & Instrumentation* 12(4):395–427.

Dirk Hovy. 2015. Demographic Factors Improve Classification Performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*. ACL, pages 752–762.

Barbara J Juhasz and Melvin J Yap. 2013. Sensory Experience Ratings for over 5,000 mono-and Disyllabic Words. *Behavior Research Methods* 45(1):160–168.

Barbara J Juhasz, Melvin J Yap, Joanna Dicke, Sarah C Taylor, and Margaret M Gullick. 2011. Tangible Words are Recognized Faster: The Grounding of Meaning in Sensory and Perceptual Systems. *The Quarterly Journal of Experimental Psychology* 64(9):1683–1691.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private Traits and Attributes are Predictable from Digital Records of Human Behavior. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 110(15):5802–5805.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition Ratings for 30,000 English Words. *Behavior Research Methods* 44(4):978–990.

Vasileios Lampos, Nikolaos Aletras, Daniel Preoţiuc-Pietro, and Trevor Cohn. 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, pages 405–413.

Ye Liu, Luming Zhang, Liqiang Nie, Yan Yan, and David S Rosenblum. 2016. Fortune Teller: Predicting your Career Path. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, pages 201–207.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research* 30:457–500.

Robert R McCrae and Paul T Costa Jr. 1997. Conceptions and Correlates of Openness to Experience. *Handbook of Personality Psychology* pages 825–847.

Robert R McCrae and Oliver P John. 1992. An Introduction to the Five-Factor Model and its Applications. *Journal of Personality* 60(2):175–215.

Albert Mehrabian and James A Russell. 1974. *An Approach to Environmental Psychology*. MIT Press.

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating Personality-aware Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. EMNLP, pages 1102–1108.

Allan Paivio. 2013. *Imagery and Verbal Processes*. Psychology Press.

Gregory Park, H.Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Michal Kosinski, David Stillwell, Lyle Ungar, and Martin Seligman. 2015. Automatic Personality Assessment through Social Media Language. *JPSP* 108:934–952.

Ellie Pavlick, Johan Bos, Malvina Nissim, Charley Beller, Benjamin Van Durme, and Chris Callison-Burch. 2015a. Adding Semantics to Data-driven Paraphrasing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*. ACL, pages 1511–1522.

Ellie Pavlick, Juri Ganitkevitch, Pushpendre Rastogi, Benjamin Van Durme, and Chris Callison-Burch. 2015b. PPDB 2.0: Better Paraphrase Ranking, Fine-grained Entailment Relations, Word Embeddings, and Style Classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*. ACL, pages 425–430.

Ellie Pavlick and Ani Nenkova. 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, pages 218–224.

Daniel Preoţiuc-Pietro, Jordan Carpenter, Salvatore Giorgi, and Lyle Ungar. 2016a. Studying the Dark Triad of Personality using Twitter Behavior. In *Proceedings of the 25th ACM Conference on Information and Knowledge Management*. CIKM, pages 761–770.

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle H Ungar. 2015a. The Role of Personality, Age and Gender in Tweeting about Mental Illnesses. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. NAACL.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015b. An Analysis of the User Occupational Class through Twitter Content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*. ACL, pages 1754–1764.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel J. Hopkins, and Lyle Ungar. 2017. Beyond Binary Labels: Political Ideology Prediction of Twitter Users. In *Proceedings of the 55th Conference of the Association for Computational Linguistics*. ACL.

Daniel Preoţiuc-Pietro, Sina Samangooei, Trevor Cohn, Nicholas Gibbins, and Mahesan Niranjan. 2012. Trendminer: An Architecture for Real Time Analysis of Social Media Text. In *Workshop on Real-Time Analysis and Mining of Social Streams (RAMSS)*. ICWSM.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015c. Studying User Income through Language, Behaviour and Affect in Social Media. *PLoS ONE* .

Daniel Preoţiuc-Pietro, Wei Xu, and Lyle Ungar. 2016b. Discovering User Attribute Stylistic Differences via Paraphrasing. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI, pages 3030–3037.

Francisco Rangel, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd Author Profiling Task at PAN 2015. In *CLEF*.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying Latent User Attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*. SMUC, pages 37–44.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and H Andrew Schwartz. 2014. Developing Age and Gender Predictive Lexica over Social Media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. EMNLP, pages 1146–1151.

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender Attribution: Tracing Stylometric Evidence beyond Topic and Genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. CONNL, pages 78–86.

Jocelyn Schock, Michael J Cortese, and Maya M Khanna. 2012. Imageability Estimates for 3,000 Disyllabic Words. *Behavior Research Methods* 44(2):374–379.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, and Lyle H Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-vocabulary Approach. *PLoS One* 8.

Michael P Toglia and William F Battig. 1978. *Handbook of Semantic Word Norms*. Lawrence Erlbaum.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring User Political Preferences from Streaming Communications. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. ACL, pages 186–196.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 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, pages 1815–1827.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas. *Behavior Research Methods* 45(4):1191–1207.

David Watson and Lee Anna Clark. 1997. Extraversion and its Positive Emotional Core. *Handbook of Personality Psychology* pages 767–793.