Journal of Research in Personality 44 (2010) 363–373

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers

Tal Yarkoni

Department of Psychology and Neuroscience, UCB 345, University of Colorado at Boulder, Boulder, CO 80309, United States

ARTICLE INFO

Article history:
Available online 8 April 2010

Keywords:
Personality
Blogging
Language
Individual differences
Internet

ABSTRACT

Previous studies have found systematic associations between personality and individual differences in word use. Such studies have typically focused on broad associations between major personality domains and aggregate word categories, potentially masking more specific associations. Here I report the results of a large-scale analysis of personality and word use in a large sample of blogs ($N = 694$). The size of the dataset enabled pervasive correlations with personality to be identified for a broad range of lexical variables, including both aggregate word categories and individual English words. The results replicated category-level findings from previous off-line studies, identified numerous novel associations at both a categorical and single-word level, and underscored the value of complementary approaches to the study of personality and word use.

1. Introduction

People differ considerably from each other in their habitual patterns of thought, feeling and action. Not surprisingly, these differences are reflected not only in what people think, feel, and do, but also in what they *say* about what they think, feel, or do. Recent studies have identified systematic associations between personality and language use in a variety of different contexts, including directed writing assignments (Hirsh & Peterson, 2009; Pennebaker & King, 1999), structured interviews (Fast & Funder, 2008) and naturalistic recordings of day-to-day speech (Mehl, Gosling, & Pennebaker, 2006). The results of such studies have confirmed and extended previous work on personality; for example, studies have consistently identified theoretically predicted correlations between the dimensions of Extraversion and Neuroticism and usage of words related to a variety of positive and negative emotion categories (Hirsh & Peterson, 2009; Lee, Kim, Seo, & Chung, 2007; Pennebaker & King, 1999).

Despite increasing interest, investigation of the relation between personality and word use is hampered by three limitations. First, most studies have focused on writing samples collected under laboratory settings or other relatively constrained contexts. Participants are typically directed to write or talk about specific topics, e.g., one's personal history and future goals (Fast & Funder, 2008; Hirsh & Peterson, 2009), a recent personal loss (Baddeley & Singer, 2008), or daily events (Pennebaker & King, 1999). It remains unclear to what extent the results of such studies generalize to less constrained real-world situations where people's personalities can influence not only *how* they write or talk about specific topics, but also *what* topics they choose to write or talk about (cf. Pennebaker, Mehl, & Niederhoffer, 2003). The power of a more naturalistic approach is demonstrated by a series of recent studies by Mehl and colleagues, who have used the Electronically Activated Recorder (Mehl & Pennebaker, 2003) to unobtrusively sample auditory snippets of participants' real-word behavior and language use (Mehl et al., 2006; Vazire & Mehl, 2008). Mehl and colleagues have identified a large number of associations between personality and language use, a number of which had not been previously documented in laboratory studies (Mehl et al., 2006).

Second, practical constraints limit the size and scope of most writing or speech samples. Virtually all studies to date have relied on writing or speech samples that include no more than a few thousand words per participant. As discussed below, such writing samples limit the types of analyses researchers can conduct, as it is generally not possible to reliably estimate usage rates for individual words, but only for aggregate categories. Moreover, data are typically gathered from participants on a small number of occasions (often just one) spanning several hours or days; such datasets cannot be used to establish whether any identified associations between personality and language remain stable over much longer periods of time (i.e., months or years), or reflect transient influences (e.g., mood).

Finally, most previous studies have modeled the relation between personality and language at a relatively broad level. With few exceptions (e.g., Fast & Funder, 2008), studies have focused on broad personality domains such as the Big Five, and have not explored relations with narrower personality dimensions. Similarly, nearly all studies have related differences in personality to

E-mail addresses: tal.yarkoni@colorado.edu, tyarkoni@gmail.com

0092-6566/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jrp.2010.04.001

predefined semantic categories containing dozens or hundreds of words rather than to individual words (Fast & Funder, 2008; Hirsh & Peterson, 2009; Lee et al., 2007; Pennebaker & King, 1999). Although the categorical approach has taught us a great deal about the relation between personality and language, it necessarily sacrifices specificity, because statistically reliable correlations between personality traits and individual words may be "washed out" when those words are averaged or summed together with many other words. Moreover, category-based approaches are necessarily limited in their capacity to discover novel and unexpected relations between personality and word use, because the categories used to predict personality are typically developed rationally and are thus constrained by prior theory and researchers' intuitions.

To address these limitations, the present study analyzed the relation between personality and language using participants for whom extremely large and topically diverse writing samples were readily accessible—namely, bloggers. Because bloggers were free to write about any topic of their choosing, and were (at the time of writing) unaware that their writing would be analyzed in relation to personality, the data provided an naturalistic window into the influence of personality on language use that could not be influenced by demand characteristics. Although a number of previous studies have used a blog-based approach (e.g., Gill, Nowson, & Oberlander, 2009; Nowson, 2006; Nowson & Oberlander, 2007), such studies have relied on much smaller sample sizes and/or writing samples (typically <100 subjects and/or <5000 words per blog), precluding consistent detection of small effects or the use of word-level analyses.[1] In contrast, the volume of blogging data available in the present study—nearly 700 blogs, containing a mean of 115,423 words each, and spanning a mean period of 23.9 months—provided adequate power to detect even relatively small effects, and enabled the relation between personality and word use to be modeled reliably not only at the level of broad semantic categories, but also at the level of individual words. Moreover, in contrast to previous studies, most participants in the present study provided scores not only for relatively broad personality domains (e.g., the Big Five), but also for lower-order personality facets. Thus, the present dataset was uniquely positioned to support large-scale analyses of highly specific associations between personality and word use.

Although the overall focus of the present study was on exploratory analysis of personality and word use, the study also had three more specific aims: first, to test whether many of the associations previously identified in offline settings would generalize to online self-expression in a blogging sample; second, to compare the utility of category-level and word-level analyses in identifying lexical correlates of personality; and third, to identify correlations with word use not only for broad traits such as the Big Five but also for lower-order facets.

2. Method

2.1. Participants and procedure

Potential participants were identified via random searches on Google's Blog Search engine (blogsearch.google.com), and by following blog author comments left on other blogs. Because the goal was to obtain as representative a sample as possible, no inclusion or exclusion criteria were used to select for particular types of blogs, save for the exclusion of blogs that were clearly developed for commercial purposes (i.e., to sell specific products). Nearly 5000 bloggers were invited to participate via email, and approximately 10–20% of emailed bloggers agreed to participate (a more precise estimate of the response rate is not possible, because participants did not indicate whether they were referred to the study via email versus other channels such as word of mouth). Note that because most participants were recruited via email, the resulting sample was not truly random: bias could arise either because some people were more likely to publish their email address on their blog (a requisite for being contacted), or because some people were more likely to respond to the invitation than others. However, such selection effects should generally deflate rather than spuriously inflate the correlations reported here, because their primary effect would be to restrict the range of distribution of some personality traits, artificially limiting the amount of personality variance available to correlate with other variables.

Bloggers who agreed to participate were directed to the experiment website, where they provided basic demographic information and filled out a personality questionnaire. The contents of participants' blogs were subsequently downloaded and parsed using a set of custom scripts written in the Ruby programming language. For technical reasons (i.e., ease of programmatic access), only blogs hosted using Google's Blogger service were included in the present analyses.

In total, the full sample contained 694 blogs (524 female; mean age = 36.2 years, range = 18–78, sd = 11.7), though the actual sample size was smaller for some analyses because not all participants provided personality data (see below). The fact that females comprised three-quarters of the sample raised the possibility that results might be disproportionately driven by one gender; however, partial correlation analysis demonstrated that controlling for gender and age had negligible effects on the results. For virtually all analyses, >90% of statistically significant correlations continued to show a significant correlation in the same direction (detailed results of the partial correlation analyses are available from the author upon request).

Because many variables had highly skewed distributions and a large proportion of zero values (e.g., in cases where many bloggers never used a given word), I followed previous recommendations to use non-parametric tests (Delucchi & Bostrom, 2004). All correlational analyses were therefore conducted using Spearman's rank correlation coefficient (ρ).

2.2. Personality measures

Participants who accessed the experiment website were given a choice between filling out a shorter 100-item personality questionnaire and a longer 315-item questionnaire. Both versions included a public domain measure of the "Big Five" dimensions of personality, the 50-item IPIP representation of the NEO-FFI (Goldberg et al., 2006). Additionally, the 315-item questionnaire included the 300-item IPIP representation of the NEO-PI-R, a broadband inventory assessing 30 different facets of personality. Thus, 83% of participants ($N = 576$) had scores for the Big Five factors of personality (Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness), and 62% ($N = 431$) had additional facet-level scores.[2]

2.3. Category-based analyses

Category-based analyses used the standard categories provided in the Linguistic Inquiry and Word Count (LIWC) 2001 program (Pennebaker, Francis, & Booth, 2001). LIWC is the most commonly

[1] A few studies have modeled the relation between personality and word *N*-grams (Nowson, 2006; Oberlander & Gill, 2006); however, because of the smaller writing samples, these studies relied primarily on fixed-effects analyses, effectively concatenating data from multiple subjects into distinct strata prior to analysis. Because this approach does not model subject as a random variable, the results of fixed-effects analyses do not generalize beyond the studied sample.

[2] Because some subjects omitted responses for some items, sample sizes varied slightly across traits. The numbers reported here reflect only the smallest *N*'s across all traits.

used language analysis program in studies investigating the relation between word use and psychological variables (for reviews, see Pennebaker & Graybeal, 2001; Pennebaker et al., 2003). The LIWC 2001 dictionary defines over 70 different categories (e.g., Negative Emotions, Sexuality, Work, Sleeping, etc.), most of which contain several dozens or hundreds of words. Detailed descriptions and definitions of the LIWC categories are reported elsewhere (Pennebaker et al., 2001). Scores for each category were computed by dividing the number of occurrences of all words within that category by the total number of words in the blog (Pennebaker et al., 2001).

The present study analyzed 66 LIWC categories, excluding only those that were non-semantic (e.g., proportion of long words) or relevant primarily to speech (e.g., non-fluencies and fillers). Previous studies have typically analyzed only a subset of LIWC categories, often due to insufficient data and/or inadequate reliability (Pennebaker & King, 1999). These concerns were not applicable in the present study, because the sheer size of the writing sample was expected to support reliable estimation even of word categories with a relatively low base rate. A split-half reliability analysis (i.e., randomly dividing each participants posts into two halves, and then correlating each category's frequency across halves for all participants) confirmed this supposition: the mean split-half correlation for the 66 categories was .81 (range = .43–.94), and only two categories (Anxiety and Inhibition) had correlations lower than .6. I therefore included all categories in the analyses.

Statistically significant correlations were identified using a threshold of $p < .05$. However, because of the large number of statistical comparisons (66 for each trait), there was an elevated risk of Type I error. To minimize this risk, interpretation of statistically significant findings was based primarily on the aggregate pattern of results with multiple categories or traits rather than on individual correlation coefficients. Additionally, the presence or absence of statistical significance is reported in key tables using a complementary False Discovery Rate criterion (FDR), which adaptively controls the false positive rate for only those associations deemed significant, rather than for all tests conducted (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001). The FDR was set to 5%. Thus, any correlations that survived the FDR correction (i.e., underlined coefficients in Table 1) had, on average, only a 5% probability of being false positives. In most cases, the FDR criterion was close to the nominal $p < .05$ rate, suggesting that there was minimal inflation of Type I error.

2.4. Word-based analyses

To produce a normalized measure of word use that could be meaningfully compared across blogs, I divided the number of times each word occurred in a given blog by the total number of word tokens used in that blog. All words were stripped of any leading or trailing punctuation prior to analysis. Although many of the LIWC categories include all words that share a particular stem, words were left unstemmed in the present study because preliminary analysis indicated that many words with the same stem had quite different patterns of correlation with personality (e.g., "love" and "lover").

Because the vast majority of English words have a frequency of less than 1 in 10,000 words, two steps were adopted in order to increase the reliability of the single-word measures. First, only the 5608 words that occurred most frequently across all blogs were analyzed[3]; second, only blogs containing 50,000 or more words were

[3] A word was included if it was in the top 5000 either in terms of raw frequency count, collapsing across all blogs (token frequency), or in terms of the number of different blogs in which it occurred at least once (a measure akin to contextual diversity; Adelman, Brown, & Quesada, 2006). This resulted in a set of 5608 words.

Table 1
Correlations between Big Five personality traits and LIWC categories.

LIWC category	*N*	*E*	*O*	*A*	*C*
Total pronouns	0.06	0.06	−0.21***	0.11**	−0.02
First person sing.	0.12**	0.01	−0.16***	0.05	0
First person plural	−0.07	0.11**	−0.1*	0.18***	0.03
First person	0.1*	0.03	−0.19***	0.08*	0.02
Second person	−0.15***	0.16***	−0.12**	0.08	0
Third person	0.02	0.04	−0.06	0.08	−0.08
Negations	0.11**	−0.05	−0.13**	−0.03	−0.17***
Assent	0.05	0.07	−0.11**	0.02	−0.09*
Articles	−0.11**	−0.04	0.2***	0.03	0.09*
Prepositions	−0.04	−0.04	0.17***	0.07	0.06
Numbers	−0.07	−0.12**	−0.08*	0.11*	0.04
Affect	0.07	0.09*	−0.12**	0.06	−0.06
Positive emotions	−0.02	0.1*	−0.15***	0.18***	0.04
Positive feelings	0.01	0.11*	−0.11**	0.14**	−0.02
Optimism	−0.08*	0.05	0	0.15***	0.16***
Negative Emotions	0.16***	0.04	0	−0.15***	−0.18***
Anxiety	0.17***	−0.03	−0.02	−0.03	−0.05
Anger	0.13**	0.03	0.03	−0.23***	−0.19***
Sadness	0.1*	0.02	−0.03	0.01	−0.11*
Cognitive Processes	0.13**	−0.06	−0.09*	−0.05	−0.11**
Causation	0.11**	−0.09*	−0.02	−0.11**	−0.12**
Insight	0.08	0	−0.08	0.01	−0.05
Discrepancy	0.13**	−0.07	−0.12**	−0.04	−0.13**
Inhibition	0.09*	−0.13**	−0.07	−0.08	−0.05
Tentative	0.12**	−0.11*	−0.06	−0.07	−0.1*
Certainty	0.13**	0.1*	−0.06	0.05	−0.1*
Sensory processes	0.05	0.09*	−0.11**	0.05	−0.1*
Seeing	−0.01	0.03	−0.04	0.09*	0.01
Hearing	0.02	0.12**	−0.08*	0.01	−0.12**
Feeling	0.1*	0.06	−0.01	0.1*	−0.05
Social processes	−0.06	0.15***	−0.14***	0.13**	−0.04
Communication	0	0.13**	−0.06	0.02	−0.07
Other references	−0.08*	0.15***	−0.14***	0.15***	−0.02
Friends	−0.08*	0.15***	−0.01	0.11**	0.06
Family	−0.07	0.09*	−0.17***	0.19***	0.05
Humans	−0.05	0.13**	−0.09*	0.07	−0.12**
Time	0.01	−0.02	−0.22***	0.12**	0.09*
Past tense Vb.	0.03	−0.01	−0.16***	0.1*	0
Present tense Vb.	0.06	−0.01	−0.16***	0	−0.06
Future Tense Vb.	−0.02	−0.06	−0.08	−0.01	−0.01
Space	−0.09*	0.02	−0.11**	0.16***	0.04
Up	−0.1*	0.09*	−0.15***	0.11**	0.09*
Down	−0.04	−0.02	−0.11**	0.11**	0.06
Inclusive	−0.02	0.09*	0.11**	0.18***	0.07
Exclusive	0.1*	−0.06	0	−0.07	−0.16***
Motion	−0.02	0.02	−0.22***	0.14***	0.04
Occupation	0.05	−0.12**	0.01	−0.04	0.06
School	0.06	−0.07	0.02	−0.01	−−0.04
Job/work	0.07	−0.08*	0.04	−0.07	0.07
Achievement	0.01	−0.09*	−0.05	0.05	0.14***
Leisure	−0.05	0.08*	−0.17***	0.15***	0.06
Home	0	0.03	−0.2***	0.19***	0.05
Sports	−0.01	0.05	−0.14***	0.06	0
TV/movies	−0.02	0.05	0.05	−0.05	−0.06
Music	−0.02	0.13**	0.04	0.08*	−0.11**
Money/finance	0.04	−0.04	−0.04	−0.11**	−0.08
Metaphysical	−0.01	0.08	0.07	−0.01	−0.08
Religion	−0.03	0.11**	0.05	0.06	−0.04
Death	0.03	0.01	0.15***	−0.13**	−0.12**
Physical states	0.03	0.14***	−0.09*	0.09*	−0.05
Body states	0.02	0.1*	−0.04	0.09*	−0.07
Sexuality	0.03	0.17***	0	0.08*	−0.06
Eating/drinking	−0.01	0.08	−0.15***	0.03	−0.04
Sleep	0.1*	0.02	−0.14***	0.11**	−0.03

(*continued on next page*)

Table 1 (*continued*)

LIWC category	*N*	*E*	*O*	*A*	*C*
Grooming	0.05	−0.01	−0.2***	0.07	−0.05
Swear words	0.11**	0.06	0.06	−0.21***	−0.14**

Underlined coefficients are statistically significant at FDR = .05. All correlations are based on a minimum *N* of 576.

* = $p < .05$.
** = $p < .01$.
*** = $p < .001$.

included in the word-level analyses ($N = 406$). To ensure that these cut-offs were sufficient for reliable estimation of word use, a split-half reliability analysis was conducted (i.e., all posts within each blog were randomly assigned to one of two halves, and the correlation between halves was then computed across all blogs). Fig. 1 displays loess-smoothed split-half correlations for the 5000 most frequent words as a function of word rank. The analysis suggested that reliability was high to moderate (>.6) for the first 2000–3000 or so words, and somewhat lower thereafter. However, even for low-ranked words, split-half correlations generally remained above .4, a level considered acceptable for present purposes given that word-level analyses focused primarily on the aggregate pattern of associations with personality rather than individual correlations. Word-level results were thresholded at $p < .001$ in order to minimize the incidence of false positives.

3. Results

3.1. Category-based analyses: Big Five traits

Category-based analyses similar to those used in previous studies (Fast & Funder, 2008; Hirsh & Peterson, 2009; Pennebaker & King, 1999) revealed robust correlations between the Big Five traits and the frequency with which bloggers used different word categories. Table 1 displays correlations between the LIWC categories and Big Five scores (color-coded correlograms of these results as well as corresponding results for the 30 lower-order facets are presented in Figs. S1–S5 in the supporting information available online). Of the 330 different correlation coefficients between the Big Five traits and the 66 LIWC categories, 145 (43.9%) were statistically significant at $p < .05$, and 49 (14.8%) were statistically significant at $p < .001$. Moreover, the results directly replicated previous findings at a rate substantially greater than chance; specifically, the present study successfully replicated 15 of 30 correlations between Big Five dimensions and LIWC categories reported by Pennebaker and King (1999), and 15 of 24 Big Five correlations reported by Hirsh and Peterson (2009).

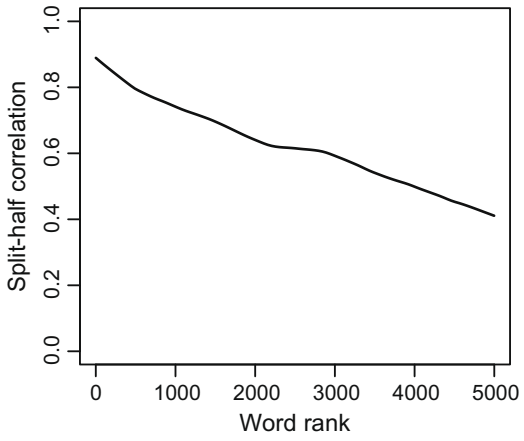

Fig. 1. Loess-smoothed plot of split-half reliability estimate as a function of word rank (ranked by frequency of occurrence in the corpus).

Importantly, many of the identified correlations converged strongly with prior findings regarding the correlates of the Big Five traits. Consistent with previous studies of personality and affective reactivity (Costa & McCrae, 1980; Larsen & Ketelaar, 1991), Neuroticism correlated positively with usage of several different negative emotion word categories, including Anxiety/Fear, Sadness, Anger, and total Negative Emotions (Table 1; Fig. S1). Conversely, Extraversion was associated with increased use of categories related to positive emotions and interpersonal interaction (Lucas & Diener, 2001; Pavot, Diener, & Fujita, 1990), including Positive Emotions, Social Processes, Friends, Sexuality, and 2nd Person References (Table 1; Fig. S2). Agreeableness, a trait characterized by an affiliative social orientation and tendency to avoid conflict with others (Graziano & Eisenberg, 1997; Graziano, Jensen-Campbell, & Hair, 1996), was positively correlated with categories indicating social communality and positive emotion (e.g., 1st Person Plural References, Family, Friends, and Positive Emotions), and negatively correlated with the use of Negative Emotion words (particularly Anger words) and Swear words.

In contrast, a number of unexpected findings were also identified. Most notably, Openness to Experience, which one might have expected to correlate positively with categories associated with emotional, intellectual, or sensory experience, was negatively correlated with 37 of the 66 LIWC categories, and positively correlated with only four categories. This pattern appeared to reflect a fundamental difference in language style rather than content (Chung & Pennebaker, 2007); people high on Openness tended to use more Articles ($\rho = .2$, $p < .001$) and Prepositions ($\rho = .17$, $p < .001$) than people low on Openness, suggesting a potential tendency to favor high-frequency function words at the expense of the lower frequency content words that made up most of the other LIWC categories.

Other unexpected findings were more specific in nature. For example, Agreeableness showed a small but statistically significant positive correlation with use of Sexual words ($\rho = .08$, $p < .05$), and Extraversion, a trait often associated with increased incentive facilitation and agentic behavior (Depue & Collins, 1999), correlated negatively with several categories reflecting goal orientation and work-related achievement (Occupation, Job/work, and Achievement; ρ's < −.08, p's < .05). One possibility is that these findings were false positives, because the analysis used a relatively liberal statistical threshold ($p < .05$, uncorrected for multiple comparisons). Alternatively, it could be that these counterintuitive findings reflected an overly broad analysis, and that more specific analyses focusing on lower-order personality facets and/or individual words rather than aggregate categories would identify more interpretable relationships. To test the latter possibility, I conducted a series of more specific facet-level and word-level analyses.

3.2. Category-based analyses: lower-order facets

Category-based analyses focusing on the 30 lower-order facets of the Big Five identified a large number of associations. Of the 1980 different correlation coefficients between the 30 facets and the 66 LIWC categories, (28.8%) were statistically significant at $p < .05$, and 152 (7.7%) were statistically significant at $p < .001$.[4] Table 3 provides a summary of the results; comprehensive results are presented in Figs. S1–S5 in the supporting information avail-

[4] Because facet-level data was only available for 62% of participants, facet-level analyses had lower power than domain-level analyses, and the reduction in the proportion of statistically significant correlations should not be taken to imply that narrower personality traits are poorer predictors of language use than broader traits.

able on-line. Not surprisingly, many of the facet-level results reaffirmed the domain-level results; for example, most Neuroticism facets correlated positively with negative emotion word use (Table 3, Fig. S1); most Extraversion facets correlated positively with use of categories related to positive emotions and social processes (Table 3, Fig. S2); and nearly all Agreeableness facets correlated negatively with the use of Anger and Swearing words.

Importantly, however, considerable facet-level heterogeneity was also identified. For each of the Big Five traits, a formal test of heterogeneity of correlated correlation coefficients (Meng, Rosenthal, & Rubin, 1992) identified at least 6 LIWC categories that showed statistically significant facet-level heterogeneity ($p < .05$). The number of heterogeneous categories was relatively low for Conscientiousness (six categories), Neuroticism (16), and Agreeableness (15). For Conscientiousness and Agreeableness, the pattern of heterogeneity could not be easily summarized (see Figs. S4 and S5); however, for Neuroticism, most of the heterogeneity appeared to stem primarily from a single facet: Self-Consciousness differed from the other facets in that it showed no positive correlation with negative affect categories, and conversely, was the only Neuroticism facet to correlate negatively with categories related to interpersonal interaction (Table 3, Fig. S1).

Extraversion and Openness showed a markedly greater degree of facet-level heterogeneity (42 and 56 heterogeneous categories, respectively). Extraversion facets displayed at least three distinct patterns of correlation with the LIWC categories (Fig. S2). First, the facets of Friendliness, Gregariousness, and Cheerfulness all showed consistent positive correlations with most categories related to positive affect, communality, and interpersonal interaction (e.g., Positive Emotions, Social Processes, 1st Person Plural references, Friends, Family, and Sexual words), whereas the other facets did not. Second, Excitement-Seeking was the only Extraversion facet to correlate positively with the Negative Emotions, Anger, and Swearing categories or negatively with the Inclusive, Home, and Grooming categories. Third, the Assertiveness and Activity Level facets showed generally weaker relations with the LIWC categories than the other facets, with the notable exception that Activity Level was the only Extraversion facet to correlate positively with categories related to goal-directed and achievement-seeking behavior (School, Job/work, and Achievement). Collectively, these findings are consistent with the notion that the affiliative and agentic aspects of Extraversion are relatively distinct and have only partially overlapping correlates (Depue & Morrone-Strupinsky, 2005; Watson & Clark, 1997).

Heterogeneity in facet-level correlations with the LIWC categories was even more striking for Openness. Most notably, the Artistic Interests and Emotionality facets showed LIWC correlations that were almost diametrically opposite to those displayed by the other four facets (Imagination, Adventurousness, Intellect, and Liberalism; Fig. S3). Artistic Interests and Emotionality correlated positively with use of the Position Emotion, Inclusive, and Physical States categories, whereas the other four facets generally showed negative correlations. Conversely, Artistic Interests and Emotionality failed to show the robust positive associations with Articles and Prepositions demonstrated by the other facets and total Openness scores. Importantly, these dissociations could not be explained by heterogeneity in the Openness facets themselves. Intercorrelations between the six facets were all directionally positive, and were moderate or strong in most cases. For example, Artistic Interests correlated .33 with Intellect, despite the fact that the two facets showed robust correlations in opposite directions with Positive Emotion word use (ρ's = .18 and −.24, respectively; p's < .0001). Thus, the facet-level analyses confirmed that the domain-level Big Five analyses reported in the previous section masked considerable and potentially important heterogeneity at the level of narrower traits.

3.3. Word-based analyses

To investigate the relation between personality and language use at the level of individual words, two sets of analyses were conducted. First, to identify the strongest word-level correlates of each personality trait, I correlated bloggers' personality scores with a set of 5068 individual words. Tables 2 and 3 summarize the results and present the top correlations for each of the Big Five traits and 30 facets, respectively (the full trait × word matrix is available on the author's website). Interestingly, there were substantial differences in the number of individual words associated with different traits. In particular, Openness correlated significantly ($p < .001$) with 393 words, whereas Neuroticism, Extraversion and Conscientiousness all correlated with fewer than 30 words.

Not surprisingly, many of the word-level associations converged with the category-level results and supported previous findings. For example, Neuroticism correlated positively with negative emotion words (e.g., 'awful', 'lazy', 'depressing', 'terrible', and 'stressful'; all ρ's $\geqslant .19$, p's < .001); Extraversion correlated positively with words reflecting social settings or experiences (e.g., 'bar', 'restaurant', 'drinking', 'dancing', 'crowd', and 'sang'; all ρ's $\geqslant .19$, p's < .001); and Openness showed strong positive correlations with words associated with intellectual or cultural experience (e.g., 'poet', 'culture', 'narrative', 'art', 'universe', and 'literature'; all ρ's $\geqslant .27$, p's < .001).

Additionally, however, the results identified numerous unanticipated correlations. Because of the large number of statistically sig-

Table 2
Top correlations between the Big Five and individual words.

Trait	No. of words sig. at $p < .001$	Top 20 words
Neuroticism	24	Awful (0.26), though (0.24), lazy (0.24), worse (0.21), depressing (0.21), irony (0.21), road (−0.2), terrible (0.2), Southern (−0.2), stressful (0.19), horrible (0.19), sort (0.19), visited (−0.19), annoying (0.19), ashamed (0.19), ground (−0.19), ban (0.18), oldest (−0.18), invited (−0.18), completed (−0.18)
Extraversion	20	Bar (0.23), other (−0.22), drinks (0.21), restaurant (0.21), dancing (0.2), restaurants (0.2), cats (−0.2), grandfather (0.2), Miami (0.2), countless (0.2), drinking (0.19), shots (0.19), computer (−0.19), girls (0.19), glorious (0.19), minor (−0.19), pool (0.18), crowd (0.18), sang (0.18), grilled (0.18)
Openness	393	Folk (0.32), humans (0.31), of (0.29), poet (0.29), art (0.29), by (0.28), universe (0.28), poetry (0.28), narrative (0.28), culture (0.28), giveaway (−0.28), century (0.28), sexual (0.27), films (0.27), novel (0.27), decades (0.27), ink (0.27), passage (0.27), literature (0.27), blues (0.26)
Agreeableness	110	Wonderful (0.28), together (0.26), visiting (0.26), morning (0.26), spring (0.25), porn (−0.25), walked (0.23), beautiful (0.23), staying (0.23), felt (0.23), cost (−0.23), share (0.23), gray (0.22), joy (0.22), afternoon (0.22), day (0.22), moments (0.22), hug (0.22), glad (0.22), fuck (−0.22)
Conscientiousness	13	Completed (0.25), adventure (0.22), stupid (−0.22), boring (−0.22), adventures (0.2), desperate (−0.2), enjoying (0.2), saying (−0.2), Hawaii (0.19), utter (−0.19), it's (−0.19), extreme (−0.19), deck (0.18)

All correlations are based on a minimum N of 331.

Table 3
Top category and word-level correlations for the lower-order facets.

Trait	No. of cats. ($p < .05$)	Top 20 LIWC categories	No. of words ($p < .001$)	Top 20 words
Neuroticism				
Anxiety	15	Feeling (0.17), Anxiety (0.16), Articles (−0.16), Space (−0.15), 1st Person Sing. (0.15), Certainty (0.13), 1st Person (0.12), Negative Emotions (0.12), Up (−0.11), Discrepancy (0.1), 2nd Person (−0.1), Affect (0.1), Negation (0.1), Grooming (0.1), Cognitive Processes (0.1)	33	Awful (0.29), sick (0.26), road (−0.26), ground (−0.25), terribly (0.25), cranky (0.25), stress (0.24), feeling (0.24), southern (−0.24), stressful (0.24), myself (0.23), though (0.23), feel (0.23), sweater (0.23), county (−0.23), scenario (0.23), ashamed (0.22), feels (0.22), oldest (−0.22), spoiled (0.22)
Anger	17	Negative Emotions (0.18), Anger (0.17), Negation (0.16), Swearing (0.14), Discrepancy (0.13), Space (−0.13), Causation (0.13), School (0.13), Cognitive Processes (0.12), Up (−0.12), 1st Person Sing. (0.11), Exclusive (0.11), Certainty (0.11), Anxiety (0.1), Feeling (0.1), Tentative (0.1), 1st Person (0.1)	14	Sick (0.24), later (−0.23), yay (0.22), road (−0.22), possibly (0.22), completely (0.21), 30 (−0.21), though (0.21), poem (−0.21), wild (−0.21), desperately (0.2), pregnancy (0.2), should not (0.2)
Depression	18	Anger (0.15), Negative Emotions (0.15), Up (−0.14), Discrepancy (0.14), Tentative (0.13), 1st Person Pl. (−0.13), Negation (0.13), Anxiety (0.12), Cognitive Processes (0.12), Articles (−0.12), Space (−0.12), Causation (0.12), Feeling (0.12), Optimism (−0.12), Swearing (0.1), 2nd Person (−0.1), Sensory Processes (0.1), Numbers (−0.09)	14	Lazy (0.24), refuse (0.23), irony (0.22), pretend (0.22), visited (−0.22), horrible (0.22), harsh (0.22), combined (−0.21), stupid (0.21), uncomfortable (0.21), though (0.21), fuck (0.2), drugs (0.2), guardian (0.2)
Self-consciousness	23	Causation (0.18), Negation (0.16), Cognitive Processes (0.16), Achievement (0.16), Tentative (0.15), Friends (−0.15), Social Processes (−0.14), Other Refs. (−0.14), 1st Person Pl. (−0.14), Occupation (0.13), Discrepancy (0.13), Communication (−0.13), Present Tense VB (0.13), Hearing (−0.12), Family (−0.12), Religion (−0.12), School (0.12), 1st Person Sing. (0.11), Exclusive (0.11), Articles (−0.1)	41	Sizes (0.27), smoke (−0.26), city (−0.25), Irish (−0.24), messy (0.24), football (−0.24), wife (−0.24), silly (0.24), street (−0.23), easier (0.23), opinions (0.23), lazy (0.23), shorter (0.23), expecting (0.23), mountain (−0.22), fit (0.22), al (−0.22), instead (0.22), realistic (0.22), fire (−0.22)
Immoderation	9	Anger (0.18), Swearing (0.16), Negative Emotions (0.14), Optimism (−0.12), 1st Person Sing. (0.12), Tentative (0.11), Negation (0.11), Articles (−0.11), 1st Person Pl. (−0.11)	3	Apart (−0.21), drops (−0.21), already (0.21)
Vulnerability	10	Feeling (0.18), Anxiety (0.16), Articles (−0.16), 1st Person Sing. (0.14), 1st Person (0.13), Causation (0.11), Discrepancy (0.11), Cognitive Processes (0.1), Grooming (0.1), 2nd Person (−0.1)	13	Lazy (0.26), awful (0.22), bull (−0.22), Southern (−0.22), al (−0.22), uncomfortable (0.22), lately (0.22), myself (0.21), though (0.21), sunset (−0.21), drop (−0.21), combined (−0.21), feeling (0.2)
Extraversion				
Friendliness	29	Friends (0.23), Leisure (0.22), 1st Person Pl. (0.22), Family (0.2), Other Refs. (0.18), Up (0.18), Social Processes (0.17), Positive Emotions (0.17), Sexual (0.16), Space (0.16), Physical States (0.15), Home (0.15), Sports (0.15), Motion (0.14), Music (0.14), Inclusive (0.14), Eating (0.14), Time (0.13), Optimism (0.13), Causation (−0.13)	47	Sang (0.27), hotel (0.26), lazy (−0.26), kissed (0.26), shots (0.26), golden (0.24), dad (0.24), girls (0.24), restaurant (0.24), eve (0.23), best (0.23), proud (0.23), miss (0.23), accept (−0.23), soccer (0.23), met (0.22), not (−0.22), brothers (0.22), interest (−0.22), cheers (0.22)
Gregariousness	30	Friends (0.26), Leisure (0.23), Sexual (0.22), Social Processes (0.2), Music (0.2), TV/Movies (0.2), Positive Emotions (0.19), Sports (0.19), Communication (0.18), Family (0.18), Positive Feelings (0.18), Humans (0.17), Articles (−0.16), Hearing (0.16), Other Refs. (0.16), Affect (0.16), 1st Person Pl. (0.14), Eating (0.14), Time (0.13), Motion (0.13)	96	Friends (0.32), girls (0.31), tickets (0.29), Friday (0.28), concert (0.27), enough (−0.27), beings (−0.27), rather (−0.27), drinks (0.27), Ryan (0.27), useful (−0.26), ticket (0.26), aka (0.26), birds (−0.25), pages (−0.25), met (0.25), gentle (−0.25), patterns (−0.25), haha (0.25), concept (−0.25)
Assertiveness	6	Communication (0.14), 2nd Person (0.11), Friends (0.11), Numbers (−0.1), Hearing (0.1), Social Processes (0.09)	7	Aka (0.27), countless (0.25), restaurants (0.23), bar (0.21), ticket (0.2), request (0.2)
Activity level	9	Time (0.15), Job/Work (0.14), Occupation (0.13), Motion (0.12), Up (0.12), Eating (0.11), Achievement (0.11), Leisure (0.11), School (0.1)	9	Contrary (−0.25), run (0.24), dolls (0.22), for. (0.22), pack (0.22), hours (0.21), 8 (0.21), fiction (−0.21), child (0.2)
Excitement-seeking	22	Anger (0.22), Swearing (0.22), Negative Emotions (0.19), Communication (0.19), Hearing (0.18), Numbers (−0.14), Grooming (−0.14), Music (0.14), Sexual (0.14), Causation (0.13), Affect (0.12), TV/Movies (0.12), Sports (0.12), Assent (0.12), Articles (−0.12), Home (−0.1), Religion (0.1), Inclusive (−0.1), 2nd Person (0.1), Present Tense VB (0.1)	72	Cats (−0.28), football (0.27), sizes (−0.27), books (−0.27), sewing (−0.26), box (−0.26), winter (−0.25), leaf (−0.25), knitting (−0.25), blankets (−0.25), delightful (−0.24), book (−0.24), piles (−0.24), I'm (0.24), haha (0.24), shelf (−0.24), asking (0.24), terrific (−0.24), gentle (−0.24), cat (−0.24)
Cheerfulness	31	Positive Emotions (0.25), Music (0.25), Positive Feelings (0.22), Affect (0.21), Friends (0.21), Sexual (0.21), 2nd Person (0.21), Leisure (0.2), Physical States (0.19), Assent (0.17), 1st Person Pl. (0.16), Other Refs. (0.16), Total Pronouns (0.16), Eating (0.15), Seeing (0.14), Social Processes (0.14), Space (0.14), Motion (0.13), Body States (0.12), 1st Person (0.12)	41	Checking (0.27), excitement (0.26), love (0.25), kidding (0.25), hot (0.25), friends (0.25), spend (0.24), shots (0.24), glory (0.23), miss (0.23), sing (0.23), girls (0.23), perfect (0.23), denied (−0.23), sweet (0.23), song (0.23), every (0.22), temporary (−0.22), dance (0.22), golden (0.22)
Openness				
Imagination	22	Home (−0.22), Time (−0.21), Death (0.19), Motion (−0.18), Up (−0.18), Family (−0.18), Past Tense VB (−0.17), Swearing (0.16), Grooming (−0.15), Leisure	105	Novel (0.29), fame (0.28), urge (0.28), decades (0.27), urban (0.27), 8th (−0.26), glance (0.26), length (0.26), poetry (0.26), literature (0.26), audience (0.26), 8 (−0.25), anniversary (−0.25),

Table 3 (*continued*)

Trait	No. of cats. ($p < .05$)	Top 20 LIWC categories	No. of words ($p < .001$)	Top 20 words
		(−0.14), Anger (0.14), Positive Emotions (−0.14), 1st Person (−0.14), Articles (0.14), Total Pronouns (−0.13), Eating (−0.13), Optimism (−0.12), 1st Person Pl. (−0.12), Social Processes (−0.12), 1st Person Sing. (−0.11)		6 (−0.25), loves (−0.25), narrative (0.25), lines (0.24), bears (0.24), thank (−0.24), humans (0.24)
Artistic interests	20	Inclusive (0.21), Positive Feelings (0.21), Music (0.2), Sexual (0.19), Seeing (0.19), Positive Emotions (0.18), Exclusive (−0.16), Leisure (0.14), Negation (−0.13), Anger (−0.13), Optimism (0.13), Inhibition (−0.12), Discrepancy (−0.12), Causation (−0.11), TV/Movies (0.11), Physical States (0.1), Cognitive Processes (−0.1), Swearing (−0.11), 1st Person Pl. (0.1), Body States (0.1)	58	Beauty (0.26), moon (0.26), blues (0.26), sky (0.26), plants (0.26), dance (0.26), beautiful (0.25), trees (0.25), planted (0.25), flowers (0.25), sang (0.25), blue (0.25), sings (0.25), danced (0.25), music (0.24), afterwards (−0.24), tree (0.24), painted (0.24), hills (0.24), outdoor (0.23)
Emotionality	22	Feeling (0.26), Sexual (0.22), Physical States (0.21), Body States (0.19), Positive Feelings (0.18), Anxiety (0.17), Affect (0.17), Sadness (0.16), 1st Person (0.16), Inclusive (0.16), 1st Person Sing. (0.16), Total Pronouns (0.15), Negative Emotions (0.14), Sleeping (0.14), Positive Emotions (0.14), Certainty (0.13), Numbers (−0.13), Friends (0.12), Assent (0.12), Humans (0.11)	36	Feel (0.29), breathe (0.29), feeling (0.28), awful (0.28), stressful (0.27), stress (0.26), fabulous (0.26), felt (0.25), heart (0.24), lucky (0.24), cried (0.23), overwhelming (0.23), sleep (0.23), hours (0.22), scared (0.22), sick (0.22), therapy (0.22), am (0.22), myself (0.22), feels (0.22)
Adventurousness	28	Grooming (−0.22), Negation (−0.21), Total Pronouns (−0.19), Present Tense VB (−0.18), 1st Person Sing. (−0.18), 1st Person (−0.18), Discrepancy (−0.15), Physical States (−0.14), Sleeping (−0.14), Home (−0.14), Affect (−0.13), Body States (−0.13), Certainty (−0.13), Cognitive Processes (−0.13), Prepositions (0.12), Tentative (−0.12), Assent (−0.12), Sensory Processes (−0.12), Exclusive (−0.12)	32	Streets (0.28), city (0.27), century (0.25), sexual (0.24), industry (0.24), businesses (0.24), south (0.23), tour (0.23), Sean (0.23), global (0.22), diaper (−0.22), immigration (0.22), countries (0.22), legal (0.22), poet (0.22), buildings (0.22), employment (0.22), west (0.21), little (−0.21), al (0.21)
Intellect	35	Eating (−0.24), Total Pronouns (−0.24), Positive Emotions (−0.24), Time (−0.24), Motion (−0.23), 1st Person (−0.23), Grooming (−0.22), 1st Person Sing. (−0.21), Articles (0.2), Physical States (−0.2), Affect (−0.2), Home (−0.2), Prepositions (0.2), Past Tense VB (−0.19), Positive Feelings (−0.18), Leisure (−0.17), Sleeping (−0.17), Sensory Processes (−0.17), 2nd Person (−0.14), Other Refs. (−0.14)	574	Against (0.37), argument (0.35), knowledge (0.35), by (0.34), sense (0.34), political (0.34), models (0.34), belief (0.34), human (0.34), historical (0.33), greater (0.33), state (0.33), universe (0.33), philosophy (0.33), humans (0.33), beings (0.33), evidence (0.32), scientists (0.32), thank (−0.32), leap (0.32)
Liberalism	36	2nd Person (−0.29), Other Refs. (−0.26), Home (−0.25), Family (−0.25), Leisure (−0.24), Positive Emotions (−0.24), Grooming (−0.24), Social Processes (−0.22), Total Pronouns (−0.21), Motion (−0.2), Sports (−0.19), Positive Feelings (−0.18), Time (−0.17), Down (−0.17), 1st Person Pl. (−0.17), Religion (−0.16), Swearing (0.16), Affect (−0.16), Prepositions (0.16), Up (−0.15)	353	Complicated (0.4), literature (0.37), particularly (0.37), prayers (−0.36), giveaway (−0.36), thankful (−0.35), hubby (−0.34), let (−0.34), unlikely (0.34), less (0.33), complex (0.33), folk (0.33), terms (0.33), fucking (0.33), entirely (0.33), structure (0.33), cultural (0.33), liberal (0.33), university (0.32), bizarre (0.32)
Agreeableness				
Trust	22	Space (0.22), Anger (−0.2), Numbers (0.2), 1st Person Pl. (0.18), Home (0.18), Leisure (0.18), Time (0.17), Motion (0.16), Up (0.16), Family (0.15), Death (−0.15), Positive Emotions (0.15), Down (0.15), Negative Emotions (−0.14), Optimism (0.13), Inclusive (0.12), Past Tense VB (0.12), Swearing (−0.11), Sports (0.11), Causation (−0.1)	56	Summer (0.31), afternoon (0.29), spent (0.27), exploring (0.27), fuck (−0.25), finishing (0.25), early (0.24), evening (0.24), Reagan (−0.24), visiting (0.24), harm (−0.23), year (0.23), drugs (−0.23), USA (−0.23), spring (0.23), two (0.23), minute (0.23), excuse (−0.23), amendment (−0.23), planned (0.23)
Morality	20	Time (0.18), Home (0.18), Swearing (−0.18), Anger (−0.16), Motion (0.15), Leisure (0.14), Family (0.14), Up (0.14), Down (0.13), Numbers (0.13), Positive Emotions (0.13), Inclusive (0.12), 1st Person Pl. (0.12), Grooming (0.11), Negative Emotions (−0.11), Space (0.11), Optimism (0.1), Death (−0.1), Past Tense VB (0.1), Total Pronouns (0.09)	44	UK (−0.26), finish (0.25), gifts (0.24), nap (0.24), finished (0.24), laundry (0.24), popcorn (0.24), day (0.23), goodness (0.23), blessed (0.23), two (0.23), guardian (−0.23), through (0.23), rest (0.23), gray (0.22), bin (−0.22), folded (0.22), sexual (−0.22), book (0.22), until (0.22)
Altruism	23	Anger (−0.18), Optimism (0.18), Leisure (0.17), 1st Person Pl. (0.16), Friends (0.16), Swearing (−0.16), Positive Emotions (0.15), Motion (0.15), Space (0.14), Family (0.14), Inclusive (0.13), Home (0.13), Up (0.13), Down (0.12), Tentative (−0.12), Other Refs. (0.11), Death (−0.1), Sports (0.1), Causation (−0.1), Time (0.1)	24	Idiot (−0.24), hug (0.24), blast (0.23), chips (0.23), greeted (0.23), minutes (0.22), rest (0.22), times (0.22), cup (0.22), beach (0.22), solved (−0.22), seconds (0.22), Olympic (0.22), stupid (−0.22), following (0.21), dinner (0.21), participants (0.21), die (−0.21), fabulous (0.21), sharing (0.21)
Cooperation	20	Anger (−0.26), Swearing (−0.26), Space (0.23), Numbers (0.2), Negative Emotions (−0.19), Down (0.15), Optimism (0.14), Inclusive (0.13), Money (−0.13), Communication (−0.13), Death (−0.13), Up (0.13), Hearing (−0.12), Prepositions (0.12), Positive Emotions (0.12), Causation (−0.11), Negation (−0.11), Motion (0.1), Home (0.1), Time (0.1)	51	Fuck (−0.3), unusual (0.3), asshole (−0.28), spring (0.27), particular (0.26), porn (−0.25), lake (0.25), paid (−0.25), seemed (0.25), two (0.25), fucking (−0.25), enemies (−0.24), sexual (−0.24), tree (0.24), four (0.24), adventure (0.24), determined (0.23), gay (−0.23), occasionally (0.23), activity (0.23)
Modesty	9	Motion (0.16), Achievement (0.14), Time (0.13), Home (0.13), Positive Emotions (0.11), Past Tense VB (0.1), Grooming (0.1), Sleeping (0.1), Family (0.09)	19	Audience (−0.25), increasingly (−0.25), decades (−0.25), doctor (0.24), recent (−0.24), toys (0.24), cities (−0.23), streets (−0.22), infection (0.22), style (−0.22), city (−0.21), crowds (−0.21), decade (−0.21), Russian (−0.21), box (0.21), involves (−0.21),

(*continued on next page*)

Table 3 (*continued*)

Trait	No. of cats. ($p < .05$)	Top 20 LIWC categories	No. of words ($p < .001$)	Top 20 words
				category (−0.21), cherry (0.21), model (−0.21)
Sympathy	6	Inclusive (0.13), Family (0.12), Anger (−0.11), Prepositions (0.11), Feeling (0.11), Swearing (−0.11)	28	Particular (0.26), since (−0.24), strength (0.24), information (0.24), assured (0.24), anyways (−0.23), require (0.23), providing (0.23), increased (0.22), courage (0.22), particularly (0.22), hoped (0.22), health (0.22), t (−0.22), em (−0.22), fascinating (0.22), conversation (0.22), ways (0.21), fewer (0.21), children (0.21)
Conscientiousness				
Self-efficacy	14	Negation (−0.13), Up (0.13), Leisure (0.13), Anger (−0.13), Prepositions (0.12), 1st Person Pl. (0.12), Discrepancy (−0.12), Optimism (0.11), Tentative (−0.11), Articles (0.11), Achievement (0.11), Negative Emotions (−0.11), Cognitive Processes (−0.1), Space (0.1)	4	Fired (0.23), Roberts (0.22), rough (−0.21), Hawaii (0.21)
Orderliness	9	Time (0.14), Anger (−0.14), Death (−0.13), Home (0.12), Grooming (0.12), 1st Person (0.12), Music (−0.11), 1st Person Sing. (0.1), Metaphysical States (−0.1)	27	Desperate (−0.27), routine (0.26), tbsp (0.26), vegetables (0.25), garlic (0.24), temperature (0.24), carrots (0.23), melted (0.23), snack (0.22), salad (0.22), popcorn (0.22), ps (−0.22), days (0.22), terror (−0.22), jail (−0.21), warm (0.21), enjoying (0.21), with (0.21), extreme (−0.21), cheese (0.21)
Dutifulness	18	Anger (−0.2), Swearing (−0.18), Time (0.16), Home (0.14), Motion (0.14), Optimism (0.14), Negative Emotions (−0.13), Up (0.13), Leisure (0.13), Down (0.12), Space (0.11), Hearing (−0.11), 1st Person Pl. (0.11), Achievement (0.11), Sports (0.1), Feeling (−0.1), Positive Emotions (0.1), Humans (−0.1)	20	Rest (0.26), fuck (−0.26), popcorn (0.24), hr (0.23), 14 (0.23), intelligent (−0.23), 4 (0.22), deck (0.22), bang (−0.22), pity (−0.22), 5 (0.22), lots (0.21), stack (0.21), 8 (0.21), 2 (0.21), finished (0.21), determine (0.21), pathetic (−0.21), visit (0.2), extreme (−0.2)
Achievement striving	19	Anger (−0.23), Negative Emotions (−0.17), Swearing (−0.16), Occupation (0.14), Exclusive (−0.14), Job/Work (0.14), Negation (−0.13), Optimism (0.12), Achievement (0.12), Death (−0.12), Tentative (−0.12), Discrepancy (−0.12), other (−0.11), Sadness (−0.11), Humans (−0.11), Music (−0.11), Metaphysical States (−0.1), Hearing (−0.1), School (0.1)	33	Stupid (−0.29), idiot (−0.26), religious (−0.25), vain (−0.25), decent (−0.25), wallet (−0.24), deny (−0.24), rarely (−0.24), bloody (−0.23), protest (−0.23), utter (−0.23), contrary (−0.22), shame (−0.22), majority (−0.22), soldiers (−0.22), drunk (−0.22), politically (−0.22), democracy (−0.22), fuck (−0.22), entirely (−0.21)
Self-discipline	20	Tentative (−0.18), Optimism (0.16), Exclusive (−0.16), Anger (−0.15), Discrepancy (−0.14), Cognitive Processes (−0.14), Negation (−0.13), Time (0.13), Up (0.13), Achievement (0.13), Swearing (−0.12), Leisure (0.12), Home (0.12), Family (0.11), Certainty (−0.11), Negative Emotions (−0.1), Motion (0.1), Down (0.1), Friends (0.1), Causation (−0.1)	26	Practical (−0.26), ready (0.25), HR (0.23), rarely (−0.23), boring (−0.23), quality (−0.23), overcome (−0.23), mom's (0.23), characters (−0.22), bay (0.22), 8 (0.22), it's (−0.22), involve (−0.21), until (0.21), completed (0.21), with (0.21), entirely (−0.21), clever (−0.21), Mexican (0.2), idea (−0.2)
Cautiousness	12	Swearing (−0.23), Anger (−0.21), Optimism (0.19), Negative Emotions (−0.17), Sexual (−0.15), Numbers (0.14), Music (−0.14), Hearing (−0.13), Communication (−0.12), Articles (0.11), Death (−0.1), Negation (−0.1)	15	Cheap (−0.23), rest (0.23), recovery (0.22), pace (0.22), challenging (0.22), addition (0.22), swear (−0.22), bar (−0.22), enjoy (0.21), anxious (0.21), fuck (−0.21), jokes (−0.21), terrific (0.21), extent (0.2), paid (−0.2)

All correlations are based on a minimum *N* of 263.

nificant correlations, I highlight only a few examples here. Unanticipated associations included correlations between Self-Consciousness and 'sizes' ($\rho = .27$), Intellect and 'against' ($\rho = .37$), Trust and 'summer' ($\rho = .31$), and Cooperation and 'unusual' ($\rho = .3$), to name a few (all of these examples survived even an extremely conservative Bonferroni correction for 5000 comparison—i.e., $p < .00001$).

Achieving a fuller understanding of these unexpected findings would require extensive contextual analysis that is beyond the scope of the present article (e.g., Manning & Schütze, 2000); however, as a cursory illustration of the potential power of such an approach, inspection of the local context of "sizes" revealed that the word was most commonly used in the context of clothing sizes (e.g., "a few sizes too big", "bras of all sizes", "dropping dress sizes", etc.), suggesting that highly self-conscious people may be more attuned and concerned with their physical appearance (interestingly, the correlation was numerically stronger for males than females; $\rho = .31$ versus .23). Thus, the exploratory word-level approach exemplified here can serve as a powerful tool for generating novel hypotheses that might be difficult to derive theoretically, and can be investigated more systematically in subsequent studies.

A second set of analyses sought to identify word-level heterogeneity within individual LIWC categories that could potentially have been masked by the category-level analyses. For each trait/category combination, I conducted a formal test of heterogeneity of correlated correlation coefficients (Meng et al., 1992). The analysis revealed significant heterogeneity for a large proportion of trait/category combinations (44%).[5] However, inspection revealed that most cases of statistically significant heterogeneity consisted of heterogeneity in the magnitude of correlation coefficients rather than their sign (i.e., coefficients were large for some words and close to zero for others). Such a pattern could potentially result solely from differences in the reliabilities of individual words (cf. Fig. 1), and therefore provided only weak evidence for true heterogeneity. I therefore focus here on three clear-cut cases in which distinct subsets of words within the same category correlated in opposite directions with personality.

First, Agreeableness correlated positively with some words in the Sexual words category (e.g., 'loves', 'love', and 'hug'; all ρ's $\geqslant .2$, p's $< .001$) but negatively with others (e.g., 'porn', 'gay', and 'fuck'; all ρ's $< -.21$, $p < .001$). This finding parsimoniously explained the counterintuitive positive correlation between Sexual words and Agreeableness alluded to earlier. The Sexual words cat-

[5] The Meng et al. (1992) test of heterogeneity is computationally intensive for large groups of variables, making it impractical to exhaustively test all trait/category combinations. The proportion reported here therefore applies only to those LIWC categories containing fewer than 100 words.

egory contained heterogeneous word subsets that referred both to love and affection as well as sexual behavior and swearwords; however, because the affection-related had much higher usage rates than the latter ones, the overall category-level scores for Sexual word use were dominated by words related to affection rather than sex. A post hoc analysis supported this supposition, as Agreeableness correlated strongly in opposite directions with two four-word categories related to love and affection ('love', 'loves', 'loved', and 'loving'; $\rho = .33$) versus sexual behavior ('fuck', 'porn', 'gay', and 'rape'; $\rho = -.4$), and the two 4-word categories were themselves negatively correlated ($\rho = -.15$, $p < .01$).

Second, Intellect (a facet of Openness reflecting interest and engagement with intellectual ideas) correlated negatively with the Space category ($\rho = -.11$, $p < .05$), despite the fact that the top 10 correlations with individual words within that category were all positive (ρ's = .18–.33, p's < .01). The explanation for this seemingly paradoxical finding was that the positively correlated words occurred relatively infrequently and predominantly expressed relational spatial concepts (e.g., 'among', 'between', 'further', and 'under'). In contrast, Intellect was negatively correlated with a number of words that denoted more concrete spatial terms ('up', 'out', 'top', 'bottom', and 'down'; ρ's = −.14 to −.16, p's < .05). Again, because the latter words were much more common, they dominated overall scores for the Space category.

Finally, Excitement-Seeking correlated heterogeneously with words within the School category. At the category-level, there was no relation between the two variables ($\rho = .02$, *ns*). However, Excitement-Seeking correlated positively with a number of sports-related words within the School category ('football', 'team', 'basketball', 'dating', and 'coach'; ρ's = .13–.27, $p < .05$), and negatively with a number of words related to academic pursuits ('books', 'book', and 'desk'; ρ's = −.13 to −.27, $p < .05$). Thus, failing to account for heterogeneity within the School category could have led to the erroneous conclusion that Excitement-Seeking was entirely unrelated to word use reflecting scholastic pursuits.

4. Discussion

Individual differences in personality have previously been linked to differences in linguistic style in laboratory and experience-sampling settings (Fast & Funder, 2008; Hirsh & Peterson, 2009; Mehl et al., 2006; Pennebaker & King, 1999). The present study replicated many of these findings in a large and heterogeneous sample of blogs, suggesting that personality exerts similar influences on offline and online forms of self-expression. The results converge with other recent findings suggesting that, contrary to popular wisdom, people do not present themselves in an idealized and overly positive way online (Turkle, 1997), and maintain online identities that reflect the way they genuinely see themselves and are seen by others (Back et al., 2010; Vazire & Gosling, 2004).

Importantly, in addition to replicating previous associations, the present findings extend previous research in several ways. First, the results address several methodological limitations of other recent studies that used data-driven approaches to investigate the relation between personality and online self-expression. For example, Nowson and Oberlander used an *N*-gram based approach to identify phrases associated with differences in the Big Five dimensions in email (Oberlander & Gill, 2006) and blog corpora (Nowson, 2006). Their results were broadly congruent with the present findings and previous off-line studies; however, the studies were underpowered (i.e., they had small *N*'s and/or writing samples), identified relatively few associations, and relied primarily on fixed-effects analyses that technically do not afford generalization of conclusion beyond the studied sample. Nowson and Oberlander (2007) analyzed a much larger sample of blogs ($N = 1672$), but used an unvalidated convenience measure of personality, and had limited writing samples (<5000 words per participant) that precluded reliable estimation of all but the most common words and phrases. In contrast, the present study had sufficient power to detect relatively modest effects for many individual English words even when modeling subject as a random variable.

Second, previous studies found relatively sparing correlations with personality (Hirsh & Peterson, 2009) or focused on restricted sets of word categories (Fast & Funder, 2008; Mehl et al., 2006; Pennebaker & King, 1999). The typically explanation for such an approach is that most word categories are not relevant to personality or are insufficiently reliable for analysis—for example, several authors have emphasized the value of studying function words rather than content words (e.g., Chung & Pennebaker, 2007; Pennebaker et al., 2003). In contrast, the present study identified multiple personality correlates for virtually all LIWC categories, suggesting that personality plays a relatively pervasive role in shaping the language people use, and that diffuse associations with both function and content words can be reliably identified given a sufficiently large dataset.

Third, correlations with personality were identified not only for relatively broad word categories, but also for individual words. The increased specificity afforded by word-level analyses can facilitate research in a number of ways. One benefit is that word-level analyses can help to identify novel associations between personality and language that can subsequently be tested more systematically (e.g., the aforementioned association between self-consciousness and 'sizes'). Another benefit is that researchers can potentially test more fine-grained hypotheses about personality. For example, rather than simply demonstrating that Neurotic individuals use more negative emotion words, the present findings suggest that Neuroticism may be associated primarily with adjectival words used to describe events in a negative way (e.g., 'awful', 'depressing', 'terrible', and 'stressful') rather than nouns connoting actual negative events. Finally, word-level analyses can help refine existing categorization schemes in a "bottom-up" manner (cf. Oberlander & Gill, 2006; Pennebaker et al., 2003). For example, in the present study, Agreeableness correlated in opposite directions with distinct subsets of words within the LIWC Sexual words categories, suggesting that category might be profitably subdivided into distinct Love and Sex categories.

Fourth, the present findings suggest that some traits are more strongly expressed in people's online writing than others. Most notably, Openness showed considerably stronger associations with both category-level and word-level language use than the other traits.[6] This finding appeared to reflect increased use of more formal language and greater discussion of a broad range of intellectual topics. Moreover, because Openness is positively correlated with a broad range of cognitive abilities, including vocabulary (Ackerman & Heggestad, 1997; Gignac, Stough, & Loukomitis, 2004), it is reasonable to suppose that highly Open individuals use "big words" more often, effectively resulting in strong positive correlations for many individual words (but conversely, producing negative correlations for LIWC categories that tend to be made up of a relatively small number of high-frequency words). A post hoc analysis confirmed this supposition, as Openness correlated robustly with the mean string length of all words used ($\rho = .26$, $p < .001$).

[6] Nowson and Oberlander (2007) previously suggested that Openness scores might be too high among bloggers (and the range consequently too restricted) to support analysis of language use patterns. Although absolute Openness scores were indeed high in the present study (mean score = 41.3 out of a possible 50), substantial variability remained (sd = 5.82), and restriction of range clearly did not prevent robust associations from emerging.

Finally, from a methodological standpoint, the present findings underscore the importance of exploring relations between personality and language at multiple levels of analysis (cf. Fast & Funder, 2008). Previous studies have tended to focus on broad personality traits such as the Big Five and/or broad categories of words (Hirsh & Peterson, 2009; Mehl et al., 2006; Oberlander & Nowson, 2006; Pennebaker & King, 1999). To my knowledge, the present study represents the first effort to systematically relate both broad and narrow personality traits to both categorical and single-word measures of language use. The results demonstrate that high-level associations between broad traits and aggregate word categories can mask, and in some cases even contradict, robust but relatively narrow associations.

The present study also had a number of limitations worth noting. First, it is likely that selection bias influenced the results to some extent, since only a small proportion of bloggers publicly display their email addresses, and of those who do, only a fraction agreed to participate when contacted via email. It is reasonable to suppose that bloggers who participated had systematically different personalities from those who did not (e.g., they might be more Agreeable or Open). Such a discrepancy could potentially bias results, and also effectively rules out direct comparison of bloggers' personalities with those of the general population (since any differences in personality cannot be attributed specifically to blogging status). However, it is important to note that the primary consequence of selection bias would be a restricted distribution of personality scores among self-selected participants, which would generally tend to *deflate* effect sizes and statistical power, leading to results that actually underestimate the magnitude of true population effects.

Second, the magnitude of many of the present correlations identified in the present study may seem relatively modest in comparison to the effect sizes reported by several previous studies. Indeed, the single largest correlation between any LIWC category and Big Five trait was .23 (Fig. S4)—a magnitude close to the *mean* statistically significant effect size found in some previous studies (Fast & Funder, 2008; Hirsh & Peterson, 2009; Mehl et al., 2006). It is important to remember, however, that effect sizes for statistically significant effects typically vary inversely with sample size, because when power is relatively low, one must capitalize on chance in order to obtain statistically significant results (Ioannidis, 2008; Yarkoni, 2009). This point can be illustrated by comparing the present results with those of a recent study by Hirsh and Peterson (2009). In the present study, 44% of all correlations between the Big Five traits and LIWC categories were statistically significant, yet the mean absolute correlation was only .14. In contrast, Hirsh and Peterson (2009) found fewer than 15% of tested effects to be statistically significant, yet obtained a much larger mean statistically significant r of .23. These seemingly paradoxical results are easily reconciled if one supposes that the true effects under investigation were actually relatively modest in *both* studies, but that Hirsh and Peterson's (2009) results, which stemmed from a much smaller sample ($N = 94$), were more susceptible to sampling error, and hence, effect size inflation (Ioannidis, 2008; Yarkoni, 2009).[7] Thus, far from flagging a problem with the present methodology, the modest effect sizes found in the present study and other large-sample studies (Pennebaker & King, 1999) are likely to be more representative of the true population effects.

Finally, although the present study explored language use at the level of both aggregate categories and individual words, all language variables were ultimately derived from simple counts of word use, and no contextual factors or higher-order semantic variables were taken into consideration. By contrast, human observers can rely on a much broader array of contextual and semantic cues when inferring other people's personalities from their writing and/or websites (e.g., Back et al., 2010; Marcus, Machilek, & Schutz, 2006; Vazire & Gosling, 2004). A human blog reader can distinguish incidental word uses from key phrases; comprehend irony and sarcasm; evaluate non-linguistic aspects of blog presentation (e.g., color selection, font size, use of images, etc.); and, in general, can develop sophisticated mental models of who a blog author is and how he or she relates to the world at large. An important challenge for future research on personality and self-expression is to determine whether more sophisticated algorithms that combine multiple channels of blog-derived information can match or exceed the accuracy displayed by human raters.

In conclusion, the present study replicated and extended previous associations between personality and language use in a uniquely large sample of blog-derived writing samples. The results underscore the importance of studying the influence of personality on word use at multiple levels of analysis, and provide a novel approach for refining existing categorical word taxonomies and identifying new and unexpected associations with personality.

Acknowledgments

This research was partially supported by NIH Award F32NR012081. The author thanks Nick Holtzman, Dave Balota, and Simine Vazire for providing valuable discussion and comments.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jrp.2010.04.001.

References

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121*(2), 219–245.

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17*(9), 814.

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., et al. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological Science, 21*(3), 372.

Baddeley, J. L., & Singer, J. A. (2008). Telling losses: Personality correlates and functions of bereavement narratives. *Journal of Research in Personality, 42*(2), 421–438.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics, 29*(4), 1165–1188.

Chung, C. K., & Pennebaker, J. W. (2007). The psychological functions of function words. In K. Fiedler (Ed.), *Social communication* (pp. 343–359). New York: Psychology Press.

Costa, P. T., & McCrae, R. R. (1980). Influence of extraversion and neuroticism on subjective well-being: Happy and unhappy people. *Journal of Personality and Social Psychology, 38*(4), 668–678.

Delucchi, K. L., & Bostrom, A. (2004). Methods for analysis of skewed data distributions in psychiatric clinical studies: Working with many zero values. *American Journal of Psychiatry, 161*(7), 1159.

Depue, R. A., & Collins, P. F. (1999). Neurobiology of the structure of personality: Dopamine, facilitation of incentive motivation, and extraversion. *Behavioral and Brain Sciences, 22*(3), 491–517 [discussion 518-469].

Depue, R. A., & Morrone-Strupinsky, J. V. (2005). A neurobehavioral model of affiliative bonding: Implications for conceptualizing a human trait of affiliation. *Behavioral and Brain Sciences, 28*(03), 313–350.

[7] In fact, the critical r value in a sample of $N = 94$ is .2, whereas it is only .08 in a sample of the present size ($N = 694$). Thus, Hirsh and Peterson (2009) would not have been able to detect the vast majority of effects identified in the present study without capitalizing on chance to some extent. Consistent with this notion, simulating 10,000 correlation tests for a population effect size of $r = .1$ reveals that the mean magnitude of statistically significant results would be .25 in a sample of $N = 94$, but only .12 in a sample of $N = 694$.

Fast, L. A., & Funder, D. C. (2008). Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. *Journal of Personality and Social Psychology, 94*(2), 334.

Gignac, G. E., Stough, C., & Loukomitis, S. (2004). Openness, intelligence, and self-report intelligence. *Intelligence, 32*(2), 133–143.

Gill, A. J., Nowson, S., Oberlander, J. (2009). *What are they blogging about? Personality, topic and motivation in blogs*. In Paper presented at the third international AAAI conference on weblogs and social media. San Jose, CA..

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., et al. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*(1), 84–96.

Graziano, W. G., & Eisenberg, N. (1997). Agreeableness: A dimension of personality. *Handbook of Personality Psychology*, 795–824.

Graziano, W. G., Jensen-Campbell, L. A., & Hair, E. C. (1996). Perceiving interpersonal conflict and reacting to it: The case for agreeableness. *Journal of Personality and Social Psychology, 70*, 820–835.

Hirsh, J. B., & Peterson, J. B. (2009). Personality and language use in self-narratives. *Journal of Research in Personality*.

Ioannidis, J. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*(5), 640.

Larsen, R. J., & Ketelaar, T. (1991). Personality and susceptibility to positive and negative emotional states. *Journal of Personality and Social Psychology, 61*(1), 132–140.

Lee, C. H., Kim, K., Seo, Y. S., & Chung, C. K. (2007). The relations between personality and language use. *The Journal of General Psychology, 134*(4), 405–413.

Lucas, R. E., & Diener, E. (2001). Understanding extraverts' enjoyment of social situations: The importance of pleasantness. *Journal of Personality and Social Psychology, 81*(2), 343–356.

Manning, C. D., & Schütze, H. (2000). *Foundations of statistical natural language processing*. MIT Press.

Marcus, B., Machilek, F., & Schutz, A. (2006). Personality in cyberspace: Personal web sites as media for personality expressions and impressions. *Journal of Personality and Social Psychology, 90*(6), 1014.

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology, 90*(5), 862.

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

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*(1), 172–175.

Nowson, S. (2006). *The language of weblogs: A study of genre and individual differences*. Edinburgh, Scotland: University of Edinburgh.

Nowson, S., & Oberlander, J. (2007). *Identifying more bloggers: Towards large scale personality classification of personal weblogs*. Paper presented at the proceedings of ICWSM.

Oberlander, J., & Nowson, S. (2006). *Whose thumb is it anyway? Classifying author personality from weblog text*. Paper presented at the COLING/ACL, Sydney, Australia.

Oberlander, J., & Gill, A. J. (2006). Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes, 42*(3), 239–270.

Pavot, W., Diener, E., & Fujita, F. (1990). Extraversion and happiness. *Personality and Individual Differences, 11*(12), 1299–1306.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: LIWC 2001*. Mahway, New Jersey: Lawrence Erlbaum Associates.

Pennebaker, J. W., & Graybeal, A. (2001). Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science, 10*(3), 90–93.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 1296–1312.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54*(1), 547–577.

Turkle, S. (1997). *Life on the screen: Identity in the age of the internet*. Touchstone Books.

Vazire, S., & Gosling, S. D. (2004). E-perceptions: Personality impressions based on personal websites. *Journal of Personality and Social Psychology, 87*, 123–132.

Vazire, S., & Mehl, M. R. (2008). Knowing me, knowing you: The accuracy and unique predictive validity of self-ratings and other-ratings of daily behavior. *Journal of Personality and Social Psychology, 95*(5), 1202–1216.

Watson, D., & Clark, L. A. (1997). Extraversion and its positive emotional core. *Handbook of Personality Psychology, 767*, 793.

Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power – Commentary on Vul et al. (2009). *Perspectives on Psychological Science, 4*(3).