The Online Social Self:
An Open Vocabulary Approach to Personality

Margaret L. Kern[1], Johannes C. Eichstaedt[1], H. Andrew Schwartz[1], Lukasz Dziurzynski[1], Lyle H. Ungar[1], David J. Stillwell[2], Michal Kosinski[2], Stephanie M. Ramones[1], Martin E. P. Seligman[1]

[1] University of Pennsylvania, [2] University of Cambridge

Author Note

Margaret L. Kern, Department of Psychology, University of Pennsylvania; Johannes C. Eichstaedt, Department of Psychology, University of Pennsylvania; H. Andrew Schwartz, Computer & Information Science, University of Pennsylvania; Lukasz Dziurzynski, Department of Psychology, University of Pennsylvania; Lyle H. Ungar, Computer & Information Science, University of Pennsylvania; David J. Stillwell, Psychometrics Centre, University of Cambridge; Michal Kosinki, Psychometrics Centre, University of Cambridge; Stephanie M. Ramones, Department of Psychology, University of Pennsylvania; Martin E. P. Seligman, Department of Psychology, University of Pennsylvania

Support for this publication was provided by the Robert Wood Johnson Foundation's Pioneer Portfolio, through the "Exploring Concepts of Positive Health" grant awarded to Martin Seligman and by the University of Pennsylvania Positive Psychology Center.

Correspondence concerning this article should be addressed to Margaret L. Kern, Department of Psychology, University of Pennsylvania, 3701 Market Street, 2nd floor, Philadelphia, PA 19104. Email: mkern@sas.upenn.edu

Final accepted version, November 2013, *Assessment*. This paper is not the copy of record and may not exactly replicate the authoritative document published in the journal. The final article is available at http://dx.doi.org/10.1037/a0035048

Abstract

Objective: We present a new open language analysis approach that identifies and visually summarizes the dominant naturally occurring words and phrases that most distinguished each Big Five personality trait. **Method:** Using millions of posts from 69,792 Facebook users, we examined the correlation of personality traits with online word usage. Our analysis method consists of feature extraction, correlational analysis, and visualization. **Results:** The distinguishing words and phrases were face valid and provide insight into processes that underlie the Big Five traits. **Conclusion:** Open-ended data driven exploration of large datasets combined with established psychological theory and measures offers new tools to further understand the human psyche.

Keywords: Computational Social Science; Big Five Personality; Linguistic Analysis; Online Studies; Interdisciplinary Research

The Online Social Self: An Open Vocabulary Approach to Personality

"oh i hate september"
~ Facebook post from an individual scoring high on neuroticism

'its that time, off to meet a friend, woohoo!!!'
~ Facebook post from an individual scoring high on extraversion

Much personality research uses small samples of participants in the undergraduate laboratory, or uses transparent questionnaires. We believe there is also value in studying personality with unobtrusive methods in its natural habitat. The explosion of online social media provides an ecologically valid vehicle for obtaining "big data" for such studies (Anderson, Fagan, Woodnutt, & Chamorro-Premuzic, 2012). We present a new open vocabulary method for studying individual differences: we systematically examine the words and phrases expressed by over 69,000 Facebook users, and examine how these words illuminate personality.

The Internet is now an environment where users actively create and process information. *Social media* refers to web-based and mobile technology that allows the creation, sharing, and discussion of user-generated content, including sharing web articles and posts, text and photograph updates on daily happenings, and the broadcasting of opinions and ideas (Kietzmann, Hermkens, McCarthy, & Silverstre, 2011). The most popular online modalities currently are blogs (personal web pages ranging from daily diaries to purposeful short articles), Twitter (a micro-blogging platform in which users post up to 140 character comments), and Facebook (a social networking service and website). In this study, we focus on Facebook.

We draw on a well-documented personality model, the five-factor model (FFM), or Big Five, with factors labeled extraversion, agreeableness, conscientiousness, neuroticism/ emotional stability, and openness to experience/ intellect. The five factors are associated with

many important life outcomes (Ozer & Benet-Martinez, 2006; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). This model has withstood much controversy and provides a theoretical framework for calibrating other constructs and new methods.

Personality characteristics are revealed through both behavior and through words and linguistic styles, such as conversations with acquaintances, friends, and strangers. A highly extraverted individual walks into a room, immediately engages in conversation, and is energized by the social interaction, while the highly introverted individual avoids the social situation altogether. Beyond behavior, the words and phrases that the actor uses influence how the observer classifies and understands him or her, often with considerable accuracy, with extrinsic characteristics (extraversion, conscientiousness, and agreeableness) being easier to identify than intrinsic characteristics (neuroticism and openness) (Funder & Sneed, 1993).

Important individual differences can be encoded as single words (Goldberg, 1993). Over a decade ago, James Pennebaker developed the software program, Linguistic Inquiry and Word Count (LIWC; Pennebaker & Francis, 1999), to count word frequencies across multiple categories (e.g., positive emotion, pronouns, work, family). The program has enabled exploration of individual differences in the frequency of words that people write or speak. Numerous studies that have used LIWC suggest that single words may be more linked to our personalities than previously thought (e.g., Chung & Pennebaker, 2007; Fast & Funder, 2008; Ireland & Mehl, in press; Pennebaker, 2002; Pennebaker, Mehl, & Niederhoffer, 2003; Tausczik & Pennebaker, 2010). For example, neuroticism relates to using more negative emotion and first person singular words, whereas extraversion relates to using more positive emotion and social words (Gill, Nowson, & Oberlander, 2009; Hirsh & Peterson, 2009; Mehl, Gosling, &

Pennebaker, 2006; Pennebaker & King, 1999; Sumner, Byers, & Shearing, 2011; Yarkoni, 2010). Individuals high in agreeableness or conscientiousness use fewer swear words (Golbeck, Robles, & Turner, 2011). Across 694 bloggers and more than 100,000 words, Yarkoni (2010) found face valid correlations between individual words and the big five traits, such as “awful”, “lazy”, and “depressing” for neuroticism. Gill (2004) concluded: “personality is indeed projected and perceived through language in a computer-mediated environment” (p. 221).

In the current study, we extend prior research by using an open vocabulary analysis to capture the Big Five personality traits at a larger level than has been done previously, with personality profiles from over 69,000 Facebook users with millions of status updates. We also go beyond the single word to define characteristic groups of words. When a person judges another individual’s personality, he or she does not think in terms of how many pronouns or affect words the person uses. The LIWC program groups words into categories (e.g., family, body, causation, past tense), which gives little indication of what it is really like to be high on neuroticism or agreeableness. So we used a new method to empirically discover the words and phrases that are most related to each of the five traits.

This new method looks at the dominant distinguishing words and phrases through an open-ended vocabulary word set that includes emoticons, misspellings, and phrases. Open-ended exploration allows identification of naturally occurring language connections that closed systems such as LIWC miss. Such exploration is particularly important in social media, where nonstandard spellings and increased use of abbreviated text (e.g., wat, 2day, u, sooooo, xxxx, mga, ttyl) are common. Better precision is obtained by using phrases rather than isolated words (e.g "sick of" versus "sick" or "cant wait" versus "cant"). Given the vast number of possible

phrases, these must be automatically identified, since it would be too cumbersome to identify all possibilities *a priori*. Further, our method generates visual representations of the words, phrases, and topics that most distinguish high versus low levels of each trait. These visualizations illustrate what it is like to score high on neuroticism, extraversion, or agreeableness, with a high degree of external validity.

Method

Participants

Facebook has become the largest online social network, with over one billion active users (Facebook.com, 2012). Facebook includes the option to add third party *applications*, which allow users to enhance their social networking experience by accessing a range of content (e.g., play games, answer questionnaires). By opting into an application, the user typically grants the application developers access to profile information such as demographics and status updates. One such application is MyPersonality, created by Kosinski and Stillwell (2011) at the University of Cambridge in 2007. The application offers various personality tests, intelligence tests, and a growing number of other scales. Participants receive feedback on, for example, how extraverted or intelligent they are compared to norms. Upon first accessing the application, participants are asked to agree to the anonymous use of their test scores for research purposes. About 40% of users have optionally allowed access to their Facebook profiles (i.e., a history of the verbal status updates posted by them on their profiles).

For the purposes of this study, we considered 71,857 English-speaking users who granted access to their status updates with a minimum of 1,000 words across their posts,[1] scores on at least one of the five personality factors, and age and gender information. Before processing the data, persons indicating that they did not speak English were removed. As the age distribution was positively skewed with many users in their twenties, we limited analyses to the middle 95% of the sample in age, resulting in a final sample of 69,792 users (62.3% female). Participants were 23.36 years old on average (SD = 8.94, range 13-65). Detailed location information was unavailable, but based upon language preferences, roughly 85% were from the U.S. or Canada, 14% were from the United Kingdom or other European English speaking countries, and 1% was from other locations globally. Participants contributed about 20 million status updates and 452 million word and phrase instances (24,530 unique language features used by at least one percent of the participants).

Measures

The MyPersonality application offers various personality measures, most prominently the Big Five personality factors based on the International Personality Item Pool (IPIP; www.ipip.ori.org; Goldberg, 1999; Goldberg et al., 2006). The IPIP NEO domains are freely available to researchers, and the items have been mapped to Costa and McCrae's (1992) NEO-PR inventory with appropriate norms established (Goldberg, 1992). Participants indicate how accurately a series of statements describe them (5-point scale, 1 = very inaccurate, 5 = very accurate). Scores are automatically compiled into the five factors (extraversion, agreeableness,

[1] A minimal word criterion was needed to reduce noise from sparse responses. The choice of the 1000 word cut-off was somewhat arbitrary. We tested 500, 1000, and 2000 word cut-offs, and correlations appeared to stabilize around 1000. Future work should test the appropriate cutoff.

conscientiousness, neuroticism, openness) and standardized, and these composite scores were used in our analyses (low neuroticism is described at times as emotional stability for consistency with the other traits). **Table 1** summarizes trait descriptives, reliabilities, and correlations.

The main Facebook page allows a person to share a brief status update with "friends". Kramer (2010) notes: "this is a self-descriptive text modality, optimized and designed to elicit updates about the self, many of which contain emotional or affective content" (p. 288). For consenting participants, status updates from January 2009 to November 2011 were automatically gathered through an Application Programming Interface (API). A random identity number linked the verbatim texts to the personality scores. Upon registering for the application, participants indicated their gender and age. Before beginning analysis, status updates were stored with an id number for the person who wrote it.

Data Analyses

Our open language analysis method consists of three parts: feature extraction, correlational analysis, and visualization. A detailed description of our full process can be found in Schwartz et al., in press, and on our website (wwbp.org).

Although few personality studies have examined associations by gender (Eaton & Funder, 2001), some evidence suggests that trait manifestation through language may differ for males and females (Fast & Funder, 2008; Mehl et al., 2006). To investigate such associations in our much larger sample and to provide insight into male versus female expressions of each trait, analyses were performed with the full sample, adjusting for age and gender, and then separately for males and females. Although Mehl and colleague's (2006) investigation of gender

differences in personality expression might provide guidance for expected differences, the study included a relatively small sample (96 people) and used a different methodology (electronically activated recordings). Thus, we consider our analyses to be exploratory and do not make specific hypotheses regarding gender differences.

Feature extraction. Words and phrases (n-grams; sequences of two or three words) are automatically separated from each message. To break the text into n-grams (i.e., tokenize status updates), we use Pott's "happyfuntokenizing" (sentiment.christopherpotts.net/code-data/happyfuntokenizing.py), adding some modifications to recognize emoticons common to Facebook text (e.g., "<3", "^_^").[2] From the tokenized text, single tokens (single words), two-token sequences (2-grams), and three-token sequences (3-grams) can be compiled. From the 2-grams and 3-grams, informative phrases (e.g., *thank you*, *merry Christmas*, *text me*) are identified and automatically selected using a point-wise mutual information criteria (i.e., the ratio of the actual rate that two words occur together to the expected rate that two words should occur together according to chance; Church & Hanks, 1990; Lin, 1998); 2-grams and 3-grams not meeting the criteria are discarded.

To focus on common language, maintain adequate power, and in line with practices by prior studies (Mehl et al., 2006; Pennebaker & King, 1999), words and phrases (i.e., single words, 2-grams, and 3-grams) are restricted to those used by at least one percent of the sample. Longer phrases could be considered, but computations become increasingly challenging (as the n-gram size increases, word combinations increase exponentially, making it difficult to count the frequency of any single n-gram), and we found that the results presented

[2] See wwbp.org/data.html for further details and the modified tokenizer, which can be run on a text file.

here already contained considerable information to explore. Words and phrases are normalized by the total number of words written by the user, and then are transformed using the Anscombe transformation (1948) to stabilize the variance.

Correlational analysis. Using an ordinary least squares linear regression framework, a linear function is fit between independent variables (i.e., words and phrases, one at a time) and the personality scores derived from the IPIP measure, adjusting for gender and age. The parameter estimate (β) indicates the strength of the relation. *P* values are used to indicate significance, but as this is an exploratory method, coefficients are only considered meaningful if the *p* value is less than a two-tailed Bonferroni-corrected value of .001 (i.e., with 24,000 features, a *p* value must be less than .001 ÷ 24,000 = 4 x 10^{-9}, to be retained).

As a test of effect robustness, we cross-validated findings by examining the percentage of overlap between older data (range 01 Jan 2009 through 20 Jul 2010; n_{posts} = 6,742,747) and newer data (range 21 Jul 2010 through 07 Nov 2011; n_{posts} = 7,924,568), splitting the data by the mean date a message was posted. We compared the top 100 most predictive words for each personality factor in the older group with the 100 top most predictive words for each personality factor in the newer group. On average, 79% of the top 100 most predictive words in group 1 were within the top 100 most predictive words in group 2. In addition, we examined the split half correlation for all words by domain, and found adequate stability (average $r_{Pearson}$ = .84; $\rho_{Spearman}$ = .91).[3]

[3] Percent overlap by domain: Extraversion positive: 79%, negative 79%; Agreeableness positive: 85%, negative 77%; Conscientiousness positive: 76%, negative: 78%; Emotional stability positive: 75%, negative: 74%; Openness positive: 83%, negative 84%. Split half correlations: Extraversion: *r* = .97, ρ = .93; Agreeableness: *r* = .92, ρ = .87; Conscientiousness: *r* = .72, ρ = .91; Emotional stability: *r* = .87, ρ = .87; Openness: *r* = .73, ρ = .96.

Visualization. A key component of our method is visualization, which helps the human mind make sense of the tens of thousands of correlations. The 100 features (words and phrases) that are most positively or negatively correlated with each outcome are combined into a word cloud, using a modified version of Wordle software (www.wordle.net/advanced). To create the visualizations, we map the correlation coefficients to a size between 10 and 110, which defines the font size for a particular feature relative to the other features in a given image. Frequency is mapped to hexadecimal encodings of color, ranging from grey to blue to red. For example, a large red word is frequently used and has a stronger correlation with the trait, whereas a small blue word is less frequent and more weakly correlated. Thus, the size of the words in our visualizations indicates the strength of the correlation between the word and personality trait, and the color is used to indicate the frequency of word use (i.e., how often it occurs in posts). Finally, this information is passed into the Wordle software, generating the final word cloud image.

Results

Personality and the Open Vocabulary Approach

Figure 1 presents the words and phrases that most distinguished each trait.[4] High neuroticism included negative words such as *depression*, *lonely,* and *kill.* High extraversion included social words and phrases such as *party, girls,* and *can't_wait*, whereas low extraversion related to isolated activities, such as *internet* and *reading*. High conscientiousness included words such as *work*, *success*, and *busy*. High openness reflected the artistic domain

[4] Correlation coefficients for the words appearing in each picture are given in online supplement Table S1. Word clouds controlled for age and gender. In a supplemental analysis, we also controlled for the other four traits; the resulting images are displayed in online supplement Figure S1.

(e.g., *soul, dreams, universe, music*), whereas low openness reflected low intellectual and cultural sophistication, with high use of shorthand language (e.g., *wat, ur, 2day*), misspellings, and reduced contractions (e.g., *dont* versus *don't*).

Although the dominant words in each word cloud generally reflected what might be expected based on decades of questionnaire-based personality research, the surrounding words suggest processes underlying each trait. For example, conscientiousness included words reflecting achievement, school, and work (e.g., *success, finals, to_work, work_tomorrow, long_*day), and activities that support relaxation and balance (e.g., *weekend, family*, *workout*, *vacation, day_off, lunch_with*) and general enjoyment (e.g., *much_fun, blessed, enjoying, wonderful)*. High emotional stability (low neuroticism) reflected positive social relationships (e.g., *team*, *game*, *success*) and activities that could build life balance (e.g., *blessed, beach, sports*). High extraversion, which has been aligned with positive emotionality (Costa & McCrae, 1980), reflected hedonic elements of well-being (e.g., *party, ;), excited*), whereas agreeableness reflected more diverse eudaimonic components of well-being (e.g., *grateful*, *wonderful*, *family*, *friends*).

Swear words were very prevalent for high neuroticism, low conscientiousness, and low agreeableness. At first pass, these categories appear indistinguishable, but distinctions appear in the words surrounding the swear words. **Figure 2** presents low agreeableness, conscientiousness, and high neuroticism with the swear words removed. Low agreeableness was characterized by aggressiveness, substance abuse, and other words reflecting a hostile approach to the world (e.g., *kill, punch, knife, drunk, i_hate, racist, idiots*). Low conscientiousness was similar to low extraversion, with computer-related words (e.g.,

pokemon, youtube, bored, 3meh0.0). Low conscientiousness was also similar to low openness, with shorthand text and emoticons (e.g., *d:, 3meh0.0, xd, ftw).* Low emotional stability was distinguished by depression, loneliness, worry, and psychosomatic symptom words (e.g., *depressed, lonely, scared, headache*). Further distinction occurs in the high end of each trait (Figure 1). For example, high agreeableness includes family and religious words; emotional stability includes sport words (e.g., *lakers, basketball, soccer*), whereas high conscientiousness includes school and work-related words.

Figure 3 displays the positive correlations for each trait, separately by gender. In general, although the frequency that words were expressed varied between genders, the words themselves were often the same. For example, whereas both women and men high in agreeableness mentioned numerous religious words, men mentioned more holidays (*thanksgiving, 4th of July, happy new year*), and women expressed more emotional words (*wonderful, blessed*) and mentioned more words reflecting gratitude (*thankful, thank you*). Differences were most apparent for emotional stability; men particularly mentioned sport-related words, whereas women high on emotionally stable mentioned more religious and gratitude words.

Personality and the Closed Vocabulary Approach

To compare our results to prior research, we replicated studies that have used the closed vocabulary LIWC lexicons. We counted word occurrence in 64 of the LIWC dictionary categories (Pennebaker & Francis, 1999), and correlated category frequencies with personality scores. Categories with personality correlations of $r = \pm.10$ or greater are summarized in **Table**

2.[5] The size and pattern were consistent with prior studies. For example, extraversion related to more positive emotion words (e.g., *happy, joyful, hope*) and more sexuality words (e.g., *condom, horny, hug*). Agreeableness, conscientiousness, and emotional stability related to fewer negative emotion words (e.g., *anxious, depressed, critical, hatred*). Openness related to greater article use (e.g., *a, a lot, an, the*) and more insight words (e.g., *complex, consider, prefer, solution*). Again, few gender differences were evident; although the strength of the correlations varied slightly for men and women, the pattern of associations were relatively the same.

Discussion

Using data from over 69,000 Facebook users, we examined relations between Big Five personality and word expression in online social media by automatically identifying the dominant distinguishing words and phrases associated with each trait. By condensing thousands of correlations visually, meaningful relations became apparent. Distinguishing words are face valid, and surrounding words provide insight into how personality traits are manifest in everyday language.

The visualizations are a core component of this technique. Rather than relying on numerical correlations between topics and personality tests that may have little real-life meaning, the visualizations highlight the dominant salient characteristics, which may bring us closer to understanding life from a person's perspective and enabling self-knowledge. Big data research is often exploratory in nature, and tens of thousands of correlations can be

[5] As r = .10 is often described as a small effect size, for simplicity we present these values. See online supplement Table S2 for full trait/category correlations for the full sample and separated by gender.

"significant" but not "meaningful". In contrast, the adage "a picture is worth a thousand words" takes on new meaning as a picture of words is a particularly appealing method. What is it like to be high in neuroticism? The word clouds paint a rather depressing picture, with sadness, loneliness, fear, and pain dominating the image.

Although different words dominate each trait, there is also considerable overlap, especially in the conscientiousness, agreeableness, and emotional stability word clouds. Digman's (1997) proposed two higher order personality factors, α and β, that underlie the Big Five factors and serve as the basis of two different theoretical systems. Factor β -- personal growth or self-actualization – combines extraversion and intellect (openness). In line with Digman's description, high levels of extraversion reflected outgoingness, expressiveness, and activity, while high levels of openness reflected creativity, imagination, and cultural sophistication. Openness to experiences has been related to social attitudes, choosing friends and spouses, political involvement, and cultural progression (McCrae & Sutin, 2009). Low openness was particularly characterized by misspellings and the use of contractions of contractions (e.g., *dont* versus *don't*), reflecting a lack of verbal sophistication.

Factor α, underlying conscientiousness, agreeableness, and emotional stability, may reflect either a social desirability factor or the socialization process itself (Digman, 1997). The word clouds again support such a higher factor. On the high end, socially acceptable activities and virtuous language were apparent, including religious type words (e.g., *the_lord, church, blessings, psalm*) and words that might build strong social relationships (e.g., *blessed, workout, basketball, team, thanksgiving*), which have been linked to good health and other desirable outcomes (e.g., McCullough, Hoyt, Larson, Koenig, & Thoresen, 2000; Pressman & Cohen, 2005;

Taylor, 2007). High agreeableness included well-being (e.g., *excited, wonderful, amazing, blessed*) and positive social relationships (e.g., *love_you_all, thank_you, friends_and_families*). High conscientiousness included physical activities (e.g., *the gym, workout, training*), spending time with family (e.g., *family, dinner with*), and a balance between work and play (e.g., *success, hard work, relaxing, much fun*), reflecting mature socialization processes (Vaillant, 2012).

On the low end, swear words and psychopathology appeared. Neuroticism has been linked to anxiety, depression, and substance use disorder (Kotov, Garmez, Schmidt, & Watson, 2010), and is evident with words such as *depressed, lonely,* and *anxiety*. A negative spiral may ensue, in which an individual scoring high on neuroticism feels depressed, spends more time online ruminating about how depressed he or she feels, and subsequently creates greater feelings of loneliness and despair. Low agreeableness reflected language that may trigger aggressive responses in others (e.g., *kill, hate*), pointing to socialization problems. Negative valence captured by the low levels of the α factor may be expressed more pathologically in social media contexts, whereas positive valence may be overly positive on the high ends of these traits. Potentially, clinicians could use the information contained in these word clouds to help identify individuals caught in a negative spiral and intervene before depression and other psychopathology builds.

Differential language can potentially be compared across different groups to consider underlying processes. For example, as others have found gender differences in word use (e.g., Fast & Funder, 2008; Mehl et al., 2006), we examined males and females separately. Highly emotionally stable men mentioned various sporting activities, whereas highly emotionally stable women included social relation words. At a more fine-grained level, for extraversion,

females mention *boys* and *girls*, whereas males mention *boys* and *girl*, without the "s". For agreeableness, *Thanksgiving* correlated for males but not females. However, few clear differences were apparent. Future research will benefit from a "differential differential language analysis" that systematically compares results of one group with another and directly tests which words most differentiate two groups on a trait.

Implications for Assessment

Gosling and colleagues (2002) suggest that people leave behavioral traces of themselves in the physical spaces that they inhabit. Similarly, our study suggests that people leave traces of themselves in the online environment. Building upon Funder's (1995) realistic accuracy model, Kluemper, Rosem, and Mossholder (2012) hypothesized that social networking sites enable a sufficient amount of information to be expressed such that others can accurately perceive the Big Five personality characteristics. Indeed, our results suggest that personality traits are reflected in natural word use, and that traits can be better understood through differential language analysis. Much can be learned about personality by studying the patterns of physical, social, and online environments in which people reside.

In terms of personality assessment, this differential language analysis technique finds the individual language that correlates with a given variable or characteristic. It can be used to suggest novel connections between behavior as manifest in writing and personality or other psychosocial variables that might not be apparent from forced answer questionnaires alone. The word clouds can help illustrate the Big Five traits, taking abstract constructs and making them concrete in terms of how personality is manifest in everyday life. Further, the method can be used as a questionnaire assessment tool; by revealing words that differentiate question or

construct responses, our technique can provide insight into what a questionnaire is actually measuring. Many self-reported measures may be face valid to the researchers, but have not been well tested in terms of how laypeople themselves understand the questions. This provides an unobtrusive method to investigate the underlying constructs that a particular measure is capturing.

Our differential language analysis process provides a novel strategy for approaching big data that combines social science theory, big data available through online social media, and tools available through computer science. Our technique challenges social sciences to think outside of the box, daring the field to use social media for assessment research. Other works might use the knowledge of which words and phrases correlate with personality factors to help in building statistical models to predict personality (for an elaboration of using penalized regression to predict personality on the basis of status updates, see Schwartz et al., in press).

Limitations

Both prior studies with LIWC and the current study found small correlations between self-reported personality and word frequency. When using individual word and phrase frequencies, most words and phrases are used at least a few times by most people, so it is unlikely that single words or phrases will relate to personality scores with an *r* larger than 0.1 or 0.2. A combination of words and phrases within one model would have larger effects.[6] Future work using machine-learning techniques can more directly address predictive models.

[6] To demonstrate, we created composite variables based on the 100 words most positively or negatively correlated with each trait. We summed the relative frequencies across all of the 100 words per user, standardized these values across our participants, and then subtracted the standardized negative composite from the standardized positive composite. We then correlated this composite variable with the personality score. Correlation coefficients

The sample size in the present study consisted of tens of thousands of individuals writing at least 1,000 words, providing high power, and thus helping the field avoid Type II errors (i.e. missing a real phenomenon). Notably, we used a very stringent criterion (i.e., requiring a language feature to be significant at a Bonferroni-corrected threshold of $p = 4 \times 10^{-9}$), and only included the 100 features most and least correlated with each trait in the word clouds, to reduce the possibility that relations are simply due to chance. Still, data mining techniques are exploratory in nature, and relations should be examined in more detail with other samples and analytic approaches.

Facebook posts, like self-report questionnaires, reflect identity and reputation management (Karl, Peluchette, & Schlaegel, 2010). We could not directly test the extent to which identity management might have occurred. However, comparisons of self-ratings, online behavior, and observer-ratings indicate that individual differences in identity management often occur in intuitively meaningful ways (Back et al., 2010; Gill, Oberlander, & Austin, 2006), such that identity management may be an important part of personality expression. Whereas participants can easily manipulate answers in transparent self-report questions, observers typically use both expressions and omissions in natural language to form personality judgments.

With such large numbers, it is easy to think that the sample is representative of the world at large. While this is a more diverse sample than undergraduate questionnaire studies, despite over one billion users (currently 15.6% of the world population and over 50% for the United States; Miniwatts Marketing Group, 2012), the sample was drawn from individuals who

---

were r = .16, .21, .25, .13, and.23 for extraversion, agreeableness, conscientiousness, neuroticism, and openness respectively, all of which were larger than any single word or phrase correlation.

chose to use a personality application and then to make their profiles available to the application. Although the popularity and ease of large Internet samples is appealing, especially with growing concerns about privacy, future research needs to carefully consider shifting bias in any online sample. In a world of quickly changing technology, the sample characteristics are also likely to change. For example, several years ago, MySpace dominated the social media culture, whereas Facebook and Twitter have since become the biggest players. Computational social science needs to be flexible and ready to shift with the tide of popular interest.

Conclusion

Mehl and colleagues (2006) noted: "in many ways, people's real-world interactions within their social environments are the very things social and personality psychologists want to know about" (p. 875). Cialdini (2009) appealed to psychologists to incorporate field-based studies, noting: "unless researchers more clearly demonstrate the value of their exploration to the wider society, support will be reduced" (p. 6). The explosion of social media and the availability of large data offer personality and social psychologists both a playground for exploration and a medium to communicate directly with the public, directly addressing Cialdini's challenge.

Our very large-scale study suggests that there are major individual differences in common word expressions that are personality-based. The typical small questionnaire studies of college undergrads cannot produce such results. The LIWC categories of single words provide a computational method for turning qualitative information from essays or online blog posts into quantitative variables that could be correlated with personality. However, the LIWC categories were manually created using a top-down approach. We have added a bottom-up

approach that automatically derives words, emoticons, misspellings, and phrases most related to personality, and allows the data to tell their own story through intuitive visualizations. In conclusion, we suggest that the marriage of computational science and psychological science may enable a better understanding of the human psyche than questionnaires alone.

References

Anderson, B., Fagan, P., Woodnutt, T., & Chamorro-Premuzic, T. (2012). Facebook psychology: Popular questions answered by research. *Psychology of Popular Media Culture, 1,* 23-37.

Anscombe, F. J. (1948). The transformation of poisson, binomial and negative-binomial data. *Biometrika, 35*, 246-254.

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, G. D. (2010). Facebook profiles reflect actual personality, not self-idealization. Psychological Science, 21, 372-374.

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

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computer Linguistics, 16*, 22–29.

Cialidini, R. B. (2009). We have to break up. *Perspectives on Psychological Science, 4*, 5-6.

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

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (Neo-PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional manual.* Odessa, FL: Psychological Assessment Resources.

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology, 73*, 1246-1256.

Eaton, L. G., & Funder, D. C. (2001). Emotional experience in daily life: Valence, variability, and rate of change. *Emotion, 1*, 413–421.

Facebook.com (2012). Fact sheet. Retrieved from http://newsroom.fb.com/content/default.aspx?NewsAreaId=22

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*, 334.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 4*, 652–670.

Funder, D. C., & Sneed, C. D. (1993). Behavioral manifestations of personality: An ecological approach to judgmental accuracy. *Journal of Personality and Social Psychology, 64*, 479–490.

Gill, A. (2004). Personality and language: The projection and perception of personality in computer-mediated communication. Doctoral dissertation, University of Edinburgh, United Kingdom. Retrieved from http://homepages.inf.ed.ac.uk/agill1/papers/GillAJ2003ThesisFinal.pdf

Gill, A.J., Nowson, S., & Oberlander, J. (2009, May). *What are they blogging about? Personality, topic and motivation in Blogs.* Proceedings of the Third International ICWSM Conference. San Jose, CA.

Gill, A. J., Oberlander, J., & Austin, E. (2006). Rating e-mail personality at zero acquaintance. *Personality and Individual Differences, 40,* 497–507.

Golbeck, J., Robles, C., & Turner, K. (2011, May). *Predicting personality with social media.* In Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems - CHI '11, Vancouver, BC, 253-262.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4,* 26-42.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, 26-34.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe*, Vol. 7 (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press.

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

Gosling, S. D., Ko, S. J., Mannarelli, T., & Morris M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology, 82*, 379–398.

Hirsh, J. B., & Peterson, J. B. (2009). Personality and language use in self-narratives. *Journal of Research in Personality, 43,* 524-527.

Ireland, M. E. & Mehl, M. R. (in press). Natural language use as a marker of personality. In T. Holtgraves (Ed.), *Oxford Handbook of Language and Social Psychology.* New York: Oxford University Press.

Karl, K., Peluchette, J., & Schlaegel, C. (2010). Who's posting Facebook faux pas? A cross-cultural examination of personality differences. *International Journal of Selection and Assessment, 18*, 174–186.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons 54*, 241-251.

Kluemper, D. H., Rosem, P. A., & Mossholder, K. W. (2012) Social networking websites, personality ratings, and the organizational context: More than meets the eye? *Journal of Applied Social Psychology, 42*, 1143-1172.

Kosinski, M. & Stillwell, D.J. (2011). myPersonality Research Wiki. *myPersonality Project.* Retrieved from http://mypersonality.org/wiki

Kotov, R., Gamez, W., Schmidt, F., & Watson, D. (2010). Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin, 136*, 768-821.

Kramer, A. D. I. (2010, April). *An unobtrusive behavioral model of "gross national happiness".* CHI '10 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, GA. Retrieved from http://dmrussell.net/CHI2010/docs/p287.pdf

Lin, D. (1998, August). *Extracting collocations from text corpora*. First Workshop on Computational Terminology, Montreal, Canada.

McCrae, R. R., & Sutin, A. R. (2009). Openness to experience. In M. R. Leary and R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior* (pp. 257-273). New York: Guilford.

McCullough, M. E., Hoyt, W. T., Larson, D. B., Koenig, H. G., & Thoresen, C. (2000). Religious involvement and mortality: A meta-analytic review. *Health Psychology, 19*, 211-222.

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*, 862.

Miniwatts Marketing Group (2012). Facebook users in the world by geographic region, 2011 Q4. Internet World Statistics. Retrieved from www.internetworldstats.com/facebook.htm

Ozer, D. J., & Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology, 57*, 401-421

Pennebaker, J.W. (2002). What our words can say about us: Toward a broader language psychology. *Psychological Science Agenda, 15*, 8-9.

Pennebaker, J. W., & Francis, M. E. (1999). Linguistic Inquiry and Word Count: LIWC. Mahwah, NJ: Erlbaum.

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*, 547–577.

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

Pressman, S. D., & Cohen, S. (2005). Does positive affect influence health? *Psychological Bulletin, 131*, 925–971.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and

cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science, 2,* 313-345.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., …, & Ungar, L. H. (in press). Personality, gender, and age in the language of social media: The open vocabulary approach. *PLOS ONE*.

Sumner, C., Byers, A., & Shearing, M. (2011, December). *Determining personality traits and privacy concerns from Facebook activity.* Black Hat Briefings Conference, Abu Dhabi, United Arab Emirates.

Tausczik, Y., & Pennebaker, J.W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29,* 24-54.

Taylor, S. E. (2007). Social support. In H. S. Friedman & R. C. Silva (Eds.), *Foundation of health psychology* (pp. 145–171). New York: Oxford University Press.

Vaillant, G. E. (2012). *Triumphs of experience: The men of the Harvard Grant Study*. Cambridge, MA: Belknap Press

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality, 44*, 363-373.

Table 1

Big five personality trait descriptives and correlations.

Trait	M	SD	α	1	2	3	4	5	6
Extraversion	-0.06	1.00	.93	--					
Agreeableness	0.02	1.00	.88	.17**	--				
Conscientiousness	-0.05	1.00	.92	.19**	.18**	--			
Emotional stability	0.14	1.04	.93	.34**	.33**	.30**	--		
Openness	0.13	0.97	.85	.13**	.06**	.04**	.06**	--	
Age	0.62	0.49	--	.01**	.05**	.05**	-.15**	-.05**	--
Gender	23.70	6.82	--	0.00	.02**	.19**	.02**	-0.01	.072**

Note. Personality traits were assessed on a 5-point Likert scale (1 = very inaccurate, 5 = very accurate). Composite scores for each factor are created and standardized by the MyPersonality application; standardized composite scores are reported.

** $p < .01$

Table 2

LIWC (Linguistic Inquiry Word Count) closed vocabulary categories: Top correlations between self-reported Big 5 personality scores and LIWC categories.

Category	Example Words	Sample	E	A	C	ES	O
Achievement	accomplish, beat, master, plan, quit	Full	.01	.05	**.13**	.09	.00
		Males	.01	.05	**.12**	.08	-.02
		Females	.02	.07	**.16**	.06	.00
Articles	a, a lot, an, the	Full	-.04	.02	.07	.06	**.13**
		Males	-.04	.00	.04	.04	**.14**
		Females	-.03	.05	.10	.02	**.12**
Body	feet, hands, skin, goose bumps, head	Full	-.01	-.09	**-.12**	-.07	.05
		Males	.01	-.08	**-.11**	-.04	.03
		Females	-.02	-.09	**-.12**	-.09	.07
Causation	makes, origin, rationale, used, why	Full	-.06	-.02	-.02	-.02	**.10**
		Males	-.06	-.02	-.02	-.01	**.10**
		Females	-.07	-.01	-.01	-.03	.09
Death	alive, bury, coffin, death, fatal, war	Full	-.08	**-.10**	**-.10**	-.04	.10
		Males	-.08	**-.10**	-.09	-.08	.08
		Females	-.08	-.08	-.10	-.07	**.11**
Family	Mother, sister, uncle, wife, pa	Full	.02	.05	.09	-.02	**-.13**
		Males	.05	.03	.06	.02	-.09
		Females	.01	.04	.09	.02	**-.14**
Filler	blah, like, oh well, you know, i mean	Full	.01	-.05	**-.12**	-.04	.05
		Males	.02	-.03	**-.11**	-.02	.02
		Females	.01	-.06	**-.13**	-.06	.07
Inclusive	add, and, both, into, open, with	Full	.04	.07	**.10**	.01	.05
		Males	.04	.04	.07	.04	.09
		Females	.04	.07	**.11**	.02	.03
Insight	accept, become, believe, know, recall	Full	-.08	.01	-.01	-.04	**.14**
		Males	-.09	.02	-.03	-.03	**.15**
		Females	-.08	.01	.01	-.05	**.12**
Negative Emotion	despair, difficult, ugh, sad, hatred	Full	-.06	**-.16**	**-.18**	**-.13**	.04
		Males	-.05	**-.14**	**-.15**	**-.12**	.01
		Females	-.06	**-.17**	**-.19**	**-.17**	.06

Category	Example Words	Sample	E	A	C	ES	O
Prepositions	For, except, over, toward, with	Full	-.02	.04	**.10**	.03	.06
		Males	-.02	.03	.08	.04	.08
		Females	-.02	.05	**.12**	.00	.04
Positive Emotion	happy, gentle, proud, humor, hugs	Full	**.13**	**.14**	**.13**	.05	-.07
		Males	**.13**	**.13**	.09	.07	-.04
		Females	**.12**	**.14**	**.13**	.09	-.08
Sensory Processes	Delicious, feel, flavor, sour, press	Full	-.03	.01	-.08	-.03	**.11**
		Males	-.02	.01	**-.10**	-.02	.09
		Females	-.04	.01	-.06	-.04	**.11**
Sexuality	pregnant, rape, lust, love, prostate	Full	**.11**	-.04	-.06	-.03	.00
		Males	**.10**	-.07	-.07	-.04	-.01
		Females	**.11**	-.03	-.06	-.02	.01
Swearing	suck, crap, butt, f**, hell	Full	.01	**-.16**	**-.13**	-.04	.02
		Males	.03	**-.14**	**-.11**	-.05	-.02
		Females	.01	**-.17**	**-.14**	**-.10**	.04
Time	anymore, autumn, presently, once	Full	.02	.07	**.11**	.03	-.05
		Males	.03	.08	.08	.06	-.04
		Females	.02	.07	**.12**	.02	-.06

Note. *N* = 69,792 users. Only categories with at least one correlations of *r* = .10 or greater are shown, and correlations of *r* = .10 or stronger are bolded. See online supplemental Table 2 for all 64 categories and full correlations. E = extraversion, A = agreeableness, C = conscientiousness, N = neuroticism, O = openness to experience.

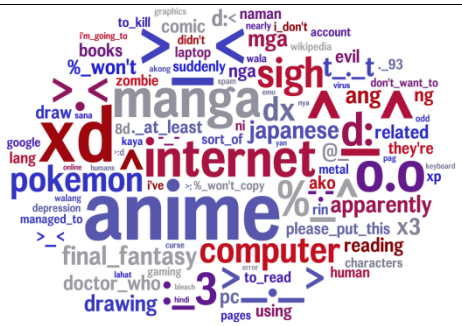
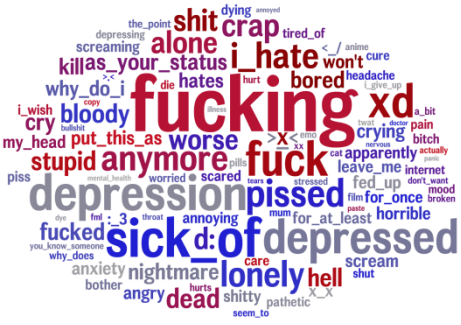
Figure Captions

Figure 1. Word clouds of the 100 words/phrases that most distinguished high (i.e., words most positively correlated with the trait) and low (i.e., words most negatively correlated with the trait) dimensions of each personality trait, adjusted for age and gender. The size of the word or phrase indicates the strength of correlation (larger = stronger) and color indicates how frequently the word or phrase appeared across user posts (dark red = frequent, grey = less frequent). Range of correlation coefficients for each image: low extraversion: *r* = -.089, -.036; high extraversion: .059, .111; low agreeableness: -.123, -.034; high agreeableness: .032, .059; low conscientiousness: -.105, -.039; high conscientiousness: .035, .069; low emotional stability: -.086, -.042; high emotional stability: .023, .047; low openness: -.090, -.039; high openness: .072, .124. Full effect size information can be found in online Supplemental Table S1.

Figure 2. Low agreeableness, conscientiousness, and emotional stability (high neuroticism), with swear words removed.

Figure 3. Male and females word clouds based on the words with the strongest positive correlations with trait scores, adjusted for age.

Figure 1

	Low	High
Extraversion		
Agreeableness		
Conscientiousness		
Emotional Stability		

Openness to Experience

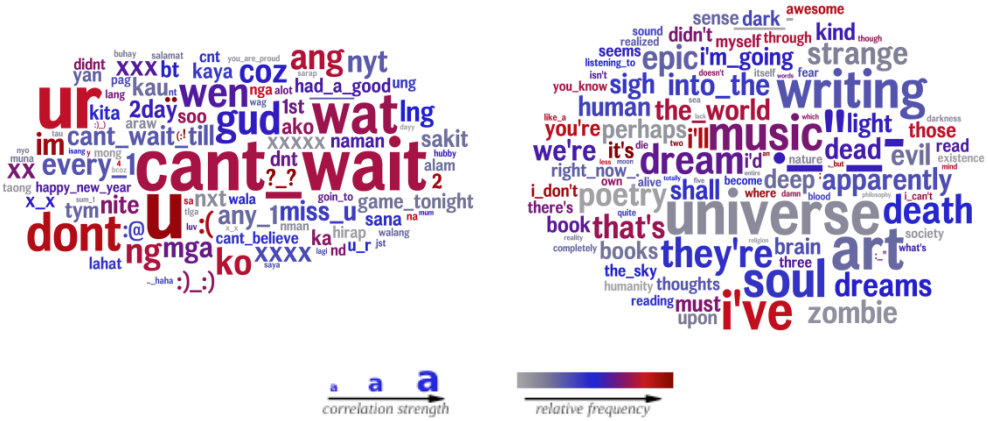


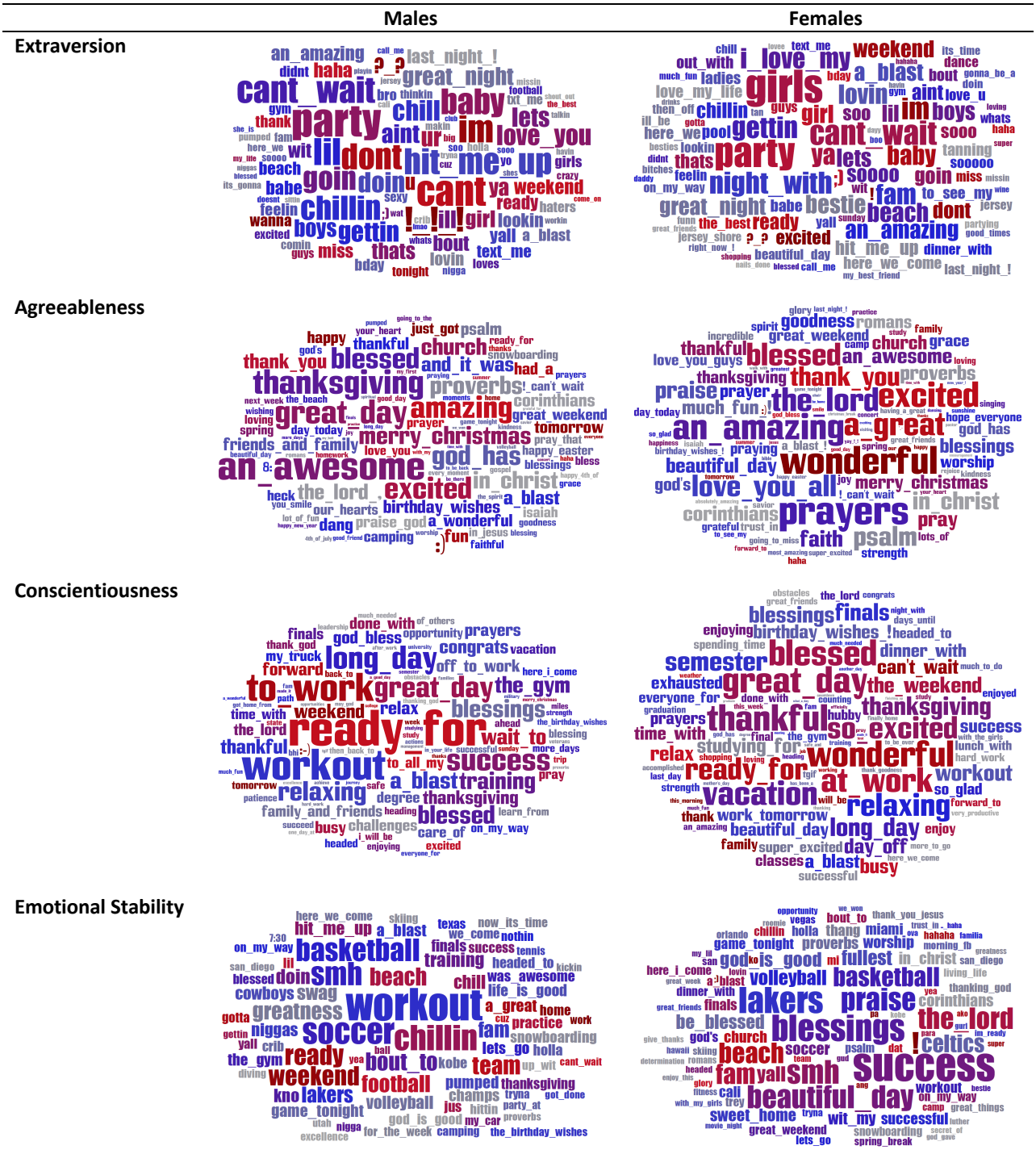
Figure 2

Low Agreeableness

Low Conscientiousness

Low Emotional Stability

Figure 3


Males
Females
Extraversion
Agreeableness
Conscientiousness
Emotional Stability

Openness to Experience