Closed and Open Vocabulary Approaches to Text Analysis: A Review, Quantitative Comparison, and Recommendations

Johannes C. Eichstaedt[1*], Margaret L. Kern[2 *], David B. Yaden[3], H. A. Schwartz[4], Salvatore Giorgi[5], Gregory Park[6], Courtney A. Hagan[5], Victoria Tobolsky[5], Laura K. Smith[5], Anneke Buffone[5], Jonathan Iwry[7], Martin E. P. Seligman[5 †], and Lyle H. Ungar[5 †]

[1] Stanford University, [2] The University of Melbourne, [4]Johns Hopkins Medicine, [4] Stony Brook University, [5] University of Pennsylvania, [6] TraitLab, Columbus, Ohio, [7] Harvard Law School
*corresponding author; [†] equal contribution

Author Note:

Johannes C. Eichstaedt, Department of Psychology & Institute for Human-Centered A.I., Stanford University, USA;
Margaret L. Kern, Melbourne Graduate School of Education, The University of Melbourne, Australia;
David B. Yaden, Department of Psychiatry and Behavioral Sciences, Johns Hopkins Medicine, USA;
H. A. Schwartz, Computer Science, Stony Brook University, USA;
Salvatore Giorgi, Department of Psychology, University of Pennsylvania, USA;
Gregory Park, Independent researcher, Columbus, Ohio;
Courtney A. Hagan, Department of Psychology, University of Pennsylvania, USA;
Anneke Buffone, Department of Psychology, University of Pennsylvania, USA;
Jonathan Iwry, Department of Psychology, University of Pennsylvania, USA;
Martin E. P. Seligman, Department of Psychology, University of Pennsylvania, USA;
Lyle H. Ungar, Computer Science Department, University of Pennsylvania, USA.

Please address correspondence to Johannes.Stanford@gmail.com, Peggy.Kern@unimelb.edu.au.

Supporting materials for this manuscript, including dictionary content summaries, comparisons between dictionaries, computer code, and topic models can be accessed at https://osf.io/h4y56

We thank Jordan Carpenter, Daniel Preoţiuc-Pietro and Shrinidhi Kowshika Lakshmikanth for their help on the project, Michal Kosinski and David J. Stillwell for providing access to the MyPersonality dataset, and Jamie Pennebaker, Ryan L. Boyd and the anonymous reviewers for their insightful comments on the manuscript.

Abstract

Technology now makes it possible to understand efficiently and at large scale how people use language to reveal their everyday thoughts, behaviors, and emotions. Written text has been analyzed through both theory-based, closed-vocabulary methods from the social sciences as well as data-driven, open-vocabulary methods from computer science, but these approaches have not been comprehensively compared. To provide guidance on best practices for automatically analyzing written text, this narrative review and quantitative synthesis compares five predominant closed- and open-vocabulary methods: Linguistic Inquiry and Word Count (LIWC), the General Inquirer, DICTION, Latent Dirichlet Allocation, and Differential Language Analysis. We compare the linguistic features associated with gender, age, and personality across the five methods using an existing dataset of Facebook status updates and self-reported survey data from 65,896 users. Results are fairly consistent across methods. The closed-vocabulary approaches efficiently summarize concepts and are helpful for understanding how people think, with LIWC 2015 yielding the strongest, most parsimonious results. Open-vocabulary approaches reveal more specific and concrete patterns across a broad range of content domains, better address ambiguous word senses, and are less prone to misinterpretation, suggesting that they are well-suited for capturing the nuances of everyday psychological processes. We detail several errors that can occur in closed-vocabulary analyses, the impact of sample size, number of words per user and number of topics included in open-vocabulary analyses, and implications of different analytical decisions. We conclude with recommendations for researchers, advocating for a complementary approach that combines closed- and open-vocabulary methods.

Non-Technical Abstract

A considerable amount of text data exists online that capture people's everyday thoughts, emotions, and behaviors. Technological advances now make it possible to analyze such data efficiently and at large scale, providing insights into everyday psychological processes as they occur in the real world. To provide guidance on best practice approaches for using such data effectively, this synthesis reviews and quantitively compares the main closed-vocabulary approaches (theoretically-derived lists of words from the social sciences) and open-vocabulary approaches (data-driven techniques from computer science that explore many words, phrases and topics) for automated text analysis. We find that the different methods are complementary; closed-vocabulary approaches provide a way to study the fundamental patterns of *how* people think and feel, whereas open-vocabulary approaches best elucidate *what* people think and feel.

Keywords: text analysis, computational social science, method comparison, language, natural language processing

Closed and Open Vocabulary Approaches to Text Analysis: A Review, Quantitative Comparison, and Recommendations

Psychological research has a long history of using a variety of methods to understand human social and psychological processes. Most of this has occurred indirectly through controlled laboratory studies, questionnaires, observations, field experiments, statistical modeling, and other approaches that attempt to mimic everyday processes. Yet it is now possible to study what people are thinking, feeling, and doing in their everyday lives, in near real time, at large scale – by analyzing the language they leave behind in digital spaces.

Humans have a long history of creating written records of their thoughts, behaviors, and experiences. Language reveals who we are, communicates information, reflects similarities and differences between groups of people, and reflects and scaffolds culture. For most of the 20th century, the rapid collection and analysis of language from tens of thousands of people was prohibitively difficult. But technological advances now make it possible to collect data on a scale that was previously inconceivable; to analyze language in principled, efficient, and replicable ways; and to identify psychological and social processes as they unfold in the real world.

In the 21st century, "those of us who use computers, and other networked devices have become a part of an emerging longitudinal, cross-sectional, and cross-cultural study" (Illiev, Dehghani, & Sagi, 2014, p. 21). This on-going real-world study encompasses large fractions of the world's population, moving far beyond the comparatively small study samples that have typified psychological studies for the past century. In particular, the mass public engagement with social media platforms such as Twitter and Facebook provide an unprecedented opportunity to study the psychological experience of millions of people – predominantly in the form of digital text.

The availability of textual data has converged with the application of computational linguistic analysis methods within the social sciences, allowing large amounts of textual data to be automatically and rapidly analyzed. Computerized text analysis was introduced in the 1960s, with various programs developed over successive decades. The original programs were **closed-vocabulary programs**, in which the researchers assign words to psychosocially relevant categories to create dictionaries, or lists of words, that are thought to represent that category (e.g., *happy*, *joy*, and *merry* are part of a ***positive emotions*** dictionary). The dictionaries have been incorporated into computer programs that allow a text to be automatically scanned, counts how often words from each dictionary occur, and then outputs the relative frequencies, which can then be used as variables in subsequent statistical analyses. Existing closed-vocabulary programs were developed within specific contexts, with specific purposes. For example, the Linguistic Inquiry and Word Count (LIWC) program was created to understand why expressive writing works (Pennebaker, Francis, & Booth, 2001)

The past two decades have introduced **open-vocabulary methods** from computer science, such as Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), word embeddings (Word2Vec; Mikolov, Chen, Corrado & Dean, 2015) and Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003). Rather than using theoretically derived categories developed from psychological and sociological theory, open-vocabulary approaches are data-driven. Algorithms identify semantically related clusters of words that naturally occur within a large set of linguistic data (see Griffiths, Steyvers, & Tenenbaum, 2007 for an excellent introduction). These clusters can then be used to predict other outcomes, gain insights about a sample, and derive new hypotheses based on patterns that appear in the data.

As of 2020, closed-vocabulary methods are the most common approach to text analysis

that have been used within psychology, with LIWC being the most popular method. Yet automated modeling has become one of the most dominant approaches to textual analysis across a number of fields, and it is only a question of time until it will become a standard tool for psychological text analysis. However, when language is modeled by computer scientists, the goal is generally to build the most accurate predictive models possible, rather than to elucidate potential psychological mechanisms or test specific theories. This difference in goals impedes the wide-spread adoption of computer science methods within the psychological sciences. Further, depending on the purpose of the study, different closed- and open-vocabulary approaches may or may not be appropriate.

Crucially, linguistic analysis methods should be judged according to the questions they are best suited to address, the insights they reveal, and the predictive power they provide. No previous review has provided a comprehensive empirical comparison of closed- and open-vocabulary approaches using the same dataset. The present comparison seeks to fill this gap and aims to serve as an introduction, orientation, and guidance to the prominent methods of text analysis for psychological science.

Here, we review the five predominant closed- and open-vocabulary approaches that have been used in the psychological literature. We trace their original purpose, emergence, and utility, and provide a quantitative comparison of these methods. While other reviews have focused on one or two approaches or have made comparisons across different datasets, here we use the same dataset to consider the ability of each approach to do the same tasks: to provide insights into psychological processes and to accurately predict individual characteristics. Supporting open science practices, we implement these analyses using a open-source language-analysis code infrastructure that is freely available. In addition, to provide guidance for the application of these methods, we test the sample sizes and words per user needed for sufficient power. For closed-vocabulary approaches, we consider drivers of prediction errors. For open-vocabulary approaches, we investigate how many topics ought to be extracted, both through a qualitative lens of conceptual nuance and through a quantitative lens of prediction accuracies.

In short, we aim to provide a comprehensive introduction and up-to-date orientation to computational methods of linguistic analysis, based on an “apples to apples” comparison on a widely-used dataset for the prominent methods since their introduction in the 1960s. While we acknowledge that predictive accuracy is generally not the goal of psychological research, our analyses provide insights into best practice approaches for effectively using the full range of available tools to understand the social and psychological processes that are revealed through people’s everyday written language.

Closed-Vocabulary Methods

Text analysis began with attempts to create a systematized approach to content analysis. Researchers developed manualized coding systems that instructed human raters how to assign codes to passages of text based on identifying "themes," which were then interpreted as the presence of a stipulated psychological construct (Mehl, 2006). Early examples include the psychoanalytical coding of the Rorschach Inkblot Test (Rorschach, 1942) and the Thematic Apperception Test (Morgan & Murray, 1935). Systematic approaches further developed through the 1960s and 70s with the growth of qualitative methodologies such as grounded theory (Glaser & Strauss, 1967). Additional qualitative coding systems have been developed over the decades (see Smith, 1992 for an overview of 14 coding systems).

Automated Text Analysis

Computers helped to automate and expedite the text analysis process. The simplest way

to quantitatively characterize a given text is to count the number of times individual words occur relative to the total number of words, ignoring word order. For example, "computational linguistic analysis is a useful psychological consideration" contains eight words, giving *"*useful*"* a relative frequency of 12.5%. Related words can be combined into **dictionaries**, or a list of words that are theoretically presumed to have something in common. For instance, the LIWC ***cognitive processes*** dictionary includes "analysis" and "consideration." A 'cognitive processes score' can be calculated by summing the relative frequencies of the words that appear in the dictionary (i.e., 25% of the words in the example above).

Dictionaries typically bring together words that the developers believe theoretically represent a particular category, similar to how items are believed to represent an underlying latent construct in a self-report measure. As such, words may not be semantically similar or commonly co-occur, but are thought to reflect explicit and implicit aspects of a construct that more holistically approximate the abstract construct when measured together. For example, Pietraszkiewicz et al.'s (2019) ***agency*** dictionary includes words such as "authoritative," "masterful," "choice," and "decide," all representing different ways that human agency might present itself. The dictionary relative frequencies can be compared across texts and correlated with other variables, using usual psychological methods of inferential statistics (Kern et al., 2016). For example, by correlating a ***social*** dictionary with gender, Newman et al. (2008) found that women tend to use more social words than men. The dictionary-based word-count approach is a seemingly transparent way to generate statistically meaningful language variables and is used by all major closed-vocabulary text analysis programs in psychology (Mehl, 2006).

To capture idiosyncrasies in how people might express the concept represented by a dictionary, most dictionaries include a generous number of synonyms. They also generally specify that different variations of the same word are counted, using wildcards that incorporate different suffixes. For example, the stem *seem** would include the word *seem*, as well as *seemed*, *seems*, *seemingly*, and *seemly*. While this aims to ensure that uses of the dictionary are detected by the program, it also means that many of the words within the dictionary are rarely or never mentioned (Alderson, 2007; Chung & Pennebaker, 2007; Pennebaker, 2011). As such, before considering the text analysis programs, we first highlight several fundamental aspects of language use that impact how these programs perform.

Statistical Fundamentals of Language Use

In language use, a few words are used much more frequently than all other words. As a minimal formal introduction, the relative frequency of words in a language follows Zipf's law (Pierce, 1980), which stipulates that the probability of encountering the *r*th most common word in a given language is inversely proportional to its rank (*r*) in that language for a normalization constant *k*:

$$P(w_r) \sim \frac{k}{r} \qquad \text{eq. 1}$$

The frequency of the *r*th most frequent word is roughly given by $P(w_r) = \frac{.1}{r}$, until about rank 1,000, such that the most common word (in English: *the*) has a probability of occurrence of $P(w_1) = .10$ (10%), followed by the words *be* (5% occurrence) and *to* (3.3% occurrence). Thus, a small set of words are very commonly used, while most words are relatively rarely used.

To illustrate, drawing on the Facebook sample used in the current review (detailed below), Figure 1 shows the frequency distribution of the 1,000 most frequent words. Even when limiting the sample to words that are used by at least 1% of the users, there remain 9,570 unique words across 258 million-word instances. However, the 96 most frequent words account for

more than 50% of word occurrences. Notably, the most common words are **function words** (***articles***, ***pronouns***, ***prepositions***, and ***conjunctions***), which fulfill mostly syntactic roles. Function words (or "style" words) have been particularly useful in psychological studies (Chung & Pennebaker, 2007; Pennebaker, 2011), providing the syntactic scaffolding of language, including ***pronouns*** (*she, I, we*), ***articles*** (*the, an, a*), ***prepositions*** (*of, as, by),* and ***conjunctions*** (*and, or, so).*

Studies find that while there are fewer than 200 common function words in the English language, they represent over half of all words used (Mehl, 2006). In contrast, **content words** are much less common, and tend to be more ideographic in nature. Accordingly, as seen in Figure 1, there are many more content words (and dictionaries to count them) but they are used much less frequently. For instance, the word *the* occurs about as frequently as all emotion words combined. Thus, function and content words have different frequency distributions: across individuals, the frequency of function words predominantly follows a normal distribution, whereas content word frequencies are predominantly highly skewed and distributed log-normally (Almodaresi et al., 2017). As a result, the frequencies of function words are often better suited than content words for analysis with standard statistical methods.

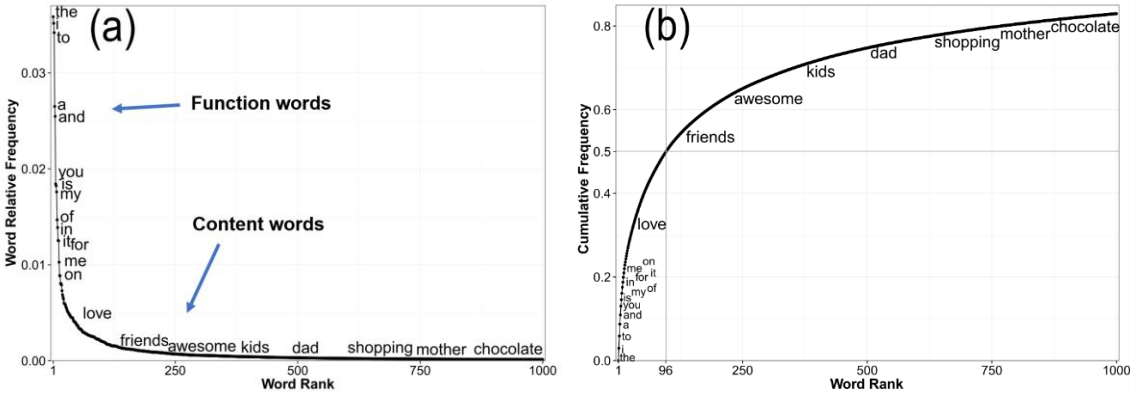


Figure 1. The relative frequency of the 1,000 most common words in a language sample of 65,896 Facebook users, shown (a) as a Zipfian distribution, in which the frequency of a word is inversely proportional to the word's frequency rank within a given language, and (b) as the cumulative frequency of the most common 1,000 words used by the sample, which account for 82% of all word occurrences. 96 words account for more than 50% of the word occurrences (marked by the cross lines in the plot).

Function words tend to be present in relatively high numbers, even in small language samples (< 500 words), making them statistically reliable markers of psychological processes that can be measured in most samples. For example, in our sample, 500 randomly selected words contained 56 pronouns, compared to 11 words expressing negative emotion. Function words are also typically used without conscious attention, thus serving as helpful markers of underlying psychological processes (Mehl, 2006). That is, one cannot typically keep track of or alter how one uses them.

All closed-vocabulary programs include both function word and content words in their dictionaries. Function word dictionaries are used more than others, for the statistical reasons review above, and function words in a mixed dictionary will be proportionally used more than other words within the dictionary. With the context of these statistical properties of language use

in mind, we turn to consideration of the most prominent closed-vocabulary programs available within psychological research.

Closed-Vocabulary Programs

Prior reviews (e.g., Neuendorf, 2002) identified 31 text analysis programs.[1] Of these, six were specifically designed to track psychological dimensions (versus providing a generic infrastructure for counting keywords) and have more than a few hundred citations in the academic literature:

- The General Inquirer (GI; Stone, Dunphy, Smith, & Ogilvie, 1966)
- DICTION (Hart, 1984)
- Linguistic Inquiry and Word Count 1993, 2001, 2007, 2015 (LIWC; Francis & Pennebaker, 1993; Pennebaker et al., 2001; Pennebaker, Booth, & Francis, 2007; Pennebaker, Boyd, Jordan, & Blackburn, 2015).
- Regressive Imagery Dictionary/Count (Martindale, 1973)
- TAS/C (Mergenthaler & Bucci, 1999)
- Gottschalk-Gleser Scales (Gleser, Gottschalk, & Sprinker, 1961; Gottchalk & Gleser, 1969)/ Psychiatric Content Analysis and Diagnosis (PCAD; Gottschalk & Bechtel, 1995, 2000)

GI, DICTION, and LIWC cover the broadest sets of content domains and are most prominent in the literature, whereas Regressive Imagery Dictionary, TAS/C, and PCAD were designed for narrow applications in clinical or psychoanalytic contexts. We thus focus on the former three programs, omitting the others from further discussion. LIWC has seemingly had the largest impact in the literature. For instance, as of April 2020, the three main versions of LIWC (2001: Pennebaker et al., 2001; 2007: Pennebaker et al., 2007; 2015: Pennebaker, et al., 2015) were cited 8,800 times. The primary citations for General Inquiry (Stone et al., 1962; Stone et al., 1966) have been cited 2,700 times. Primary references for DICTION (Hart, 1984; 2000; 2001) have been cited 280 times. We review these three programs in historical order.

The General Inquirer. GI was developed at Harvard University in the 1960s for general multi-purpose text analysis, but could also conduct analyses using custom dictionaries (Stone et al., 1962). While users were cautioned against having "unrealistic expectations" about the ease of use on mainframe computers (Kelly & Stone, 1975, p. 112), the program set the standard for the computerized programs that followed.

Considerable resources were invested in the construction of the dictionaries, with more than 10,000 human-rated annotations collected for the 12 Stanford Political Dictionaries alone (Stone et al., 1966). Between 1962 and 1965, over 25 dictionaries were developed, with additional dictionaries developed over subsequent decades. The latest version includes 182 dictionaries (see Supplementary Materials for a full list and dictionaries) matching 8,281 unique words,[2] split into three main sets: 63 Lasswell dictionaries, 107 Harvard Psychosociological dictionaries, and 12 Stanford Political dictionaries (Inquirer Home Page, 2002).

The **Lasswell dictionaries** were designed to measure eight value domains stipulated by

[1] ACTORS, CATPAC, CONCORD, Concordance 3.3, Count, CPTA, Diction 7.0, DIMAP-4, General Inquirer, Hamlet, IDENT, Intext 4.1 (now TextQuest 4.2), Lexa, LIWC, MCCALite, MECA, MonoConc, ParaConc, PCAD 2000, PROTAN, SALT, SWIFT, TABARI, TAS/C, TextAnalyst, TEXTPACK, TextSmart, The Yoshikoder, VBPro, WordStat 6.1.

[2] When determining the number of words contained within a set of dictionaries, we counted relevant word stems (e.g., for *happ**, we included *happy*, *happier* and *happiness*). Words can appear in multiple dictionaries.

Lasswell and Kaplan's (1950) influential book on power and society, and included four ***deference*** categories (***power***, ***rectitude***, ***respect***, ***affection***) and four ***welfare*** categories (***wealth***, ***well-being***, ***enlightenment***, ***skill***; Lasswell & Namenwirth, 1969). Each of these eight categories was further divided into three dictionaries: ***participants***, ***transactions*** (i.e., social allocation, or processes pertaining to the social distribution of values), and ***other***, along with a ***total*** dictionary (Weber, 1984, 1990). For example, the ***wealth-participants*** dictionary includes the words *company*, *bank*, and *customer*; the ***wealth-transactions*** dictionary includes *spend*, *bought*, and *raise*, and the ***wealth-other*** dictionary includes *car*, *own*, and *money*. Additional dictionaries were later added to cover other processes not covered by Lasswell's theory.

The **Harvard psychosociological dictionaries** were designed to extract information relevant to the leading psychological (e.g., Morgan & Murray, 1935; Murray, 1938, 1943) and sociological (e.g., McClelland, 1961) theories of the day. This set of dictionaries has undergone several updates, with the most recent form containing 107 dictionaries, such as ***virtues*** and ***feelings***, ***overstatement***, ***rituals***, ***social*** and ***cognitive*** categories, and ***motivation***-related words.

The **Stanford political dictionaries** were designed to explore the assertion that decision-making can be measured along three dimensions: ***evaluation*** (positive/negative), ***potency*** (strong/weak), and ***activity*** (active/passive) (Osgood, 1963; Osgood et al., 1957). The Stanford dictionaries sought to be comprehensive, and covered 98% of the words encountered in texts of the time (Stone et al., 1966). The dictionaries resulted from very resource-intensive annotation; multiple human judges rated every word along one, two, or three of these dimensions (e.g., *calm* = positive affect + weak + passive). This dictionary set has been used to evaluate political interactions, including some pivotal moments of geopolitical importance (e.g., Holsti, Brody, & North, 1964).

DICTION. DICTION was developed in the 1980s to analyze the "verbal tone" in 500 US presidential speeches (Hart, 1984). DICTION assumed that political texts could be characterized according to five master variables – ***activity***, ***certainty***, ***commonality***, ***optimism***, and ***realism*** – such that "if only five questions could be asked of a given passage, these five would provide the most robust understanding" (Hart, 2001, p. 45). In its current form (Version 5.0), DICTION includes 31 non-overlapping dictionaries, matching 8,578 unique words, as well as four variables that encode relative lengths of words (***complexity***), ratio of adjectives to verbs (***embellishment***), relative frequency of words repeated more than three times out of every 500 words (***insistence***), and the ratio of unique to total words (***variety***). These 35 language variables are then combined into the five master variables by adding and subtracting their standardized scores from one another (see Supplement for details). For example, ***certainty*** is derived by adding the standardized scores of *tenacity*, *leveling*, *collectives*, and *insistence*, and by subtracting *numerical terms*, *ambivalence*, *self-reference*, and *variety*. DICTION includes norm scores, which were developed from various texts, and the master variable scores of a given text can be compared to these norms. Importantly, DICTION was specifically developed for use in specific political and business contexts, such that words such as "left" or "right" were intended to refer to political leaning rather than direction. Dictionaries such as Loughran and McDonald's (2011) ***financial sentiment*** capture how positive and negative affect are understood in a business context, rather than capturing affect more broadly.

Linguistic Inquiry and Word Count. LIWC and its dictionaries were first designed in the 1990s to analyze essays written during expressive writing interventions (Francis & Pennebaker, 1992, 1993; Tausczik & Pennebaker, 2010). The program has subsequently been updated several times and has been applied to texts across a variety of domains. LIWC

dictionaries are organized hierarchically, with some dictionaries subsuming others. For instance, the ***affective processes*** dictionary is broken into ***positive emotion*** and ***negative emotion*** dictionaries, which in turn comprises ***sadness***, ***anxiety***, and ***anger*** dictionaries. As a result, when sub-dictionaries (like ***sadness***) correlate with an outcome, higher order dictionaries (like ***affective processes***) often also correlate with the outcome.

One of LIWC's biggest contributions to the literature rest on the distinction between function and content words (Chung & Pennebaker, 2007) discussed above. While GI includes multiple function word dictionaries, it was primarily the LIWC-based studies that established the importance of the function/content distinction. LIWC has revealed the importance of pronouns in revealing several different psychological processes, such increased use of first person singular "I" pronouns tracking lower status in dyadic interactions (e.g., Campbell & Pennebaker, 2003; Chung and Pennebaker, 2007; Pennebaker, 2011).

LIWC2007 has been used the most extensively in psychology. In the current review, we use the updated 2015 version, comparing LIWC2007 and LIWC2015 as a supplemental analysis. LIWC2015 provides a convenient user interface for analyzing texts. It includes 73 dictionaries, containing around 6,500 unique words (some with wildcards). LIWC's output also provides 20 summary variables, including word count and metrics based on combinations of dictionary frequencies that the creators of LIWC deemed useful (such as emotional tone).

Open-Vocabulary Methods

While automatic text analysis in psychology were first developed through closed-vocabulary approaches, open-vocabulary methods are emerging as a data-driven alternative. Among these, "clustering" approaches are of particular interest due to their capacity for reducing thousands of words into more manageable sets of variables. Specifically, one of the key advantages of these approaches is that they change the statistical representation of language from a high dimensional spaces of sparse vectors (with many zero entries, as most words do not occur in most documents) to a low dimensional space of dense vectors (often around 300 dimensions, typically all non-zero). These make them better suited as features in predictive models across a variety of tasks in Natural Language Processing and sometimes provide interpretable abstractions of language in the form of word groups (or topics).

Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) have received the most attention in the psychological literature. As of 2017, vector semantic approaches have also begun to receive attention (e.g., Bhatia, 2017; Parrigon, Woo, Tay & Wang, 2017). We briefly introduce these approaches below, in addition to Differential Language Analysis (DLA), an exploratory technique for identifying and visualizing linguistic correlates that most distinguish an outcome (Schwartz et al., 2013b).

Latent Semantic Analysis

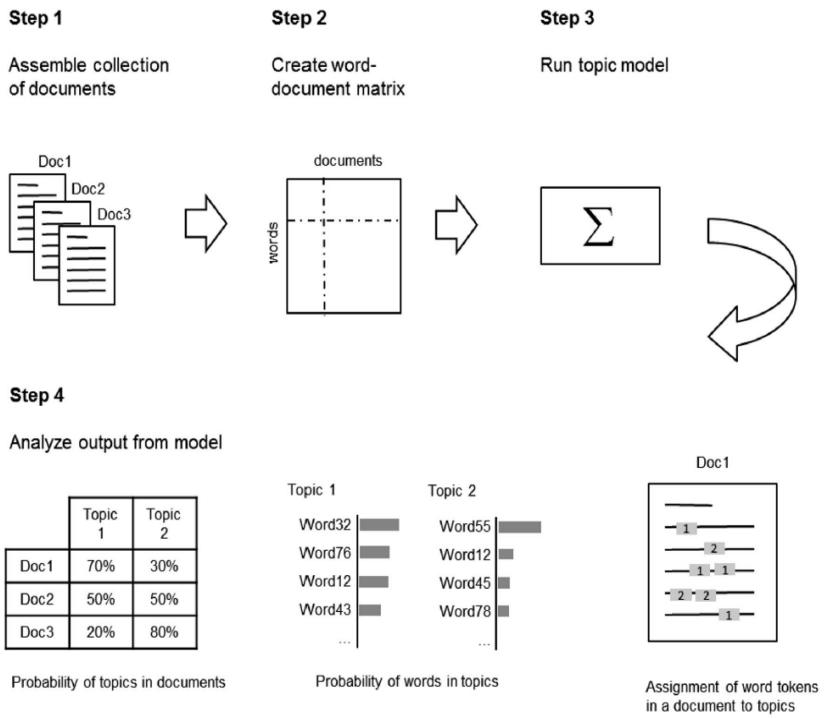
LSA was first developed in the late 1980s to determine the similarity between two bodies of text (Deerwester et al., 1988; Deerwester et al., 1990). It is similar to factor analysis, in which items are identified that align along a single dimension within a multidimensional space, resulting in a smaller number of latent factors. Factor analysis of scale items yields each participant's responses as a combination of factor scores, with survey items loading on latent factors. Similarly, LSA clusters items into latent factors (typically around 300), but in this case, the items are individual words, and the latent factors are merely a latent multi-dimensional space where by each word is represented as a point in that space. Words that are close to one another in the space tend to co-occur with the same words in documents, and thus tend to be related. (see Landauer & Dumais, 1997 for a full description and review of LSA).

Further, this dimensional representation allows LSA to quantify the semantic distance between two words as the distance between the two vectors of the words. A common metric for this distance is cosine similarity -- a normalized dot product between the two vectors capturing their similarity in vector angles and generally the extent to which the two words' contexts overlap, adjusting for baseline differences in word count. That is, it projects the vectors onto one another in the 300-dimensional space. For example, student responses on an exam can be automatically scored by calculating the distance of their response from an ideal response in the semantic space (e.g., Wolfe & Goldman, 2003). However, although LSA offers a robust method to quantify semantic differences between documents, the interpretability of its dimensions are limited. Words that negatively load on a factor are hard to interpret, and words loading onto the same factor are often not semantically coherent. This shortcoming is partly a result of approximating language as a global geometric space, which ignores the reality that most words have multiple word senses. For example, *buckle*, *belt*, and *asteroid* may cluster together, as both *buckle* and *asteroid* are semantically close to *belt*, but *buckle* is not close to *asteroid* (see Griffiths et al., 2007). In short, LSA imposes mathematical constraints that the semantic structure of language often does not follow, limiting its application for psychological language analysis. As such, we exclude LSA in our comparison.

Latent Dirichlet Allocation

LDA is a generative probabilistic clustering approach that groups words into **topics**, or coherent sets of words that cluster together across a corpus of text (Blei et al., 2003; see Griffiths et al., 2007 for an excellent review). Topics are essentially like micro-dictionaries in the closed-vocabulary approach, but the topics are generated from the data, rather than from the words that researchers believe theoretically represent that category. Like LSA, LDA is a factor analysis-type technique, which identifies latent semantic factors based on words that co-occur, but it overcomes LSA's constraints. As illustrated in Figure 2, the algorithm assumes that each word occurrence can be attributed to one or more topics generated from the corpus.

The number of topics is assigned *a priori* (this choice is non-trivial, which we consider further below). Words are assigned to a topic based on co-occurrence with other words across the corpus, and repeated until an optimal equilibrium is reached (i.e., when all of the words in the document are assigned to a set of topics with other semantically similar words). This results in a set of posterior probability distributions, which approximates the likelihood of each word occurring within each topic. These topics thus represent semantically coherent clusters of words, in which words are assigned weights based on their contribution to the topic.



	Topic 1	Topic 2
Doc1	70%	30%
Doc2	50%	50%
Doc3	20%	80%

Figure 2. The process of topic modeling using Latent Dirichlet Allocation. (1) Documents are collected and (2) represented as a word-document matric (WDM). (3) Topic models are run on the WDM. (4) The probability of topics in documents and probability of words in topics are then fit simultaneously, based on assigning individual word occurrences in documents to topics. Figure adapted from Griffiths et al., 2007.

Unlike LSA, LDA topics tend to be more semantically coherent and overcome word sense ambiguities. Through a more structured representation, LDA separates different word senses by the context in which they occur, deciding for each word which topic is most appropriate. For instance, *belt* may appear with *asteroid* in a topic together with *Jupiter*, due to co-occurring in a set of documents, whereas a separate topic would combine *belt* with *buckle* and *pants*. Additionally, word frequency is not problematic, and the confusion over how a word is used does not occur.

Topic modeling works better with a large set of documents. Importantly, the generation of topics (topic modeling) and the application (topic extraction) of previously modeled topics are two different processes that do not need to be based on the same dataset; one set of data can be used to develop the topics, and then the topics can be applied to a second dataset.[3] Thus, a large corpus can be used to *model* topics of high quality and semantic coherence, which can then be *applied* to a smaller corpus, effectively leveraging the larger dataset for building the variables and leveraging the smaller dataset to study individual characteristics.

Word embeddings

Similar to LDA topics, distributional semantic approaches (also referred to as "word embeddings" or "vector space semantics") seek to discover the different contexts in which words

[3] For example, see http://wwbp.org/data for a set of 2,000 topics modeled across 14 million Facebook statuses and then used in a variety of Twitter and Facebook datasets across a number of studies.

occur, and use these contexts (embeddings) to describe words in a low dimensional dense vector space (with typically around 300 dimensions – much fewer than the 10,000+ dimensions needed to represent if a word occurs or not). Vector semantic approaches are fundamentally based on the distributional hypothesis which states that "words that occur in similar contexts tend to have similar meanings" (Jurafsky & Martin, 2019).

LSA employed dimensionality reduction to a global word-by-document matrix, each row of which captures the frequency with which words occur in a given document (such as a diary entry, a Facebook status update, or a speech). This original matrix is the size of the number documents and number of words. The reduced version is only a fraction of that size. Word embeddings approaches (such as Word2Vec [Mikolov, Chen, Corrado & Dean, 2015] and GloVe [Pennington, Socher & Manning, 2014]) follow a different approach than direct dimensionality reduction. Instead they turn the embedding problem into a prediction problem and try to optimize a vector such that it can be used within a predictive model (e.g. a logistic regression classifier) to predict which words are in the context -- typically all words within 3 to 6 words on either side of the target word being embedded. (Mikolev, Sutskever, Chen, Corrado & Dean, 2013; Jurafsky & Martin, 2019). Thus, a sequence of words is turned into a set of prediction tasks, in which the words that actually occur are the ground truth to the classification model.[4]

For Word2Vec, the model thus learns which words are likely to occur next to each other, and this information is captured in the embeddings. Once these embeddings have been learned, a word is thus represented simply as its low dimensional vector (e.g. 300 real-valued numbers; hence, "Word2Vec"). Importantly, these vector representations can be learned on massive text data sets (even larger than those for LDA because the computational process is less intensive), and then become fixed vector representations which can be extracted from smaller study datasets. This has been the key to the success of these approaches -- they have been pre-trained on massive corpora spanning gigabytes of text data (with word counts in the 10s or 100s of billions, across vocabularies of 300 million words and phrases) which capture a large variety of nuanced language contexts by groups with access to the largest computational resources, such as Google Research (e.g., Pennington, Socher & Manning, 2014; Kenton & Toutanova, 2019).

Similarly to LSA before it, the distance between the vectors of two words in the embedding space captures semantic similarity of those words. In psychological application, Bhatia (2017) has demonstrated that these semantic distances predict the association between concepts observed across a variety of judgement tasks. Specifically, the semantic distances appear to capture the associations human judges rely on intuitively when making likelihood estimations based on "availability heuristics" the closer the concepts, the more "associated" they appear intuitively (see Bhatia, 2017 for a full discussion). As another example, Parrigon, Woo, Tay and Wang (2017) clustered the semantic distances between the vector representations of adjectives describing situations to find support for a 7-dimensional taxonomy of situations. Thus, it appears that embeddings recover regularities in our mental and physical worlds which are encoded in natural language.

In addition, the embedding vectors have proven very useful across a variety of NLP tasks. Instead of starting with raw word information, words are converted to their vectors which are used as inputs to traditional supervised models (Support Vector Machines; Random Forests; Ridge Regression) or deep learning systems. As an example, the differences ("offsets") between

[4] This general idea of trying to predict missing words, so-called "self-supervised learning," dominates to this day in how the state-of-the-art word embeddings are trained -- although the statistical models used have changed considerably (e.g., BERT; Kenton & Toutanova, 2019).

vector embeddings can capture analogous relations between words, such as that the vector for "king" minus the vector for "man" plus that for "woman" ends up providing a vector close to that of "queen" (Jurafsky & Martin, 2019). Word embeddings (and now contextual word embeddings) have become the defacto input for most natural language processing systems.

Contextual word embeddings

The word embeddings discussed in the previous section are "fixed" – that is, once they have been learned, when they are applied (or "extracted"): every word occurrence is mapped onto the same fixed list of real numbers. This vector is essentially presumed to somehow represent all of the potential roles the word could play without knowing the exact context it is being used for the application. It will no doubt often contain information irrelevant to the current context (e.g. consider the word "bank" which should capture the idea of a financial institution but being used in the sentence, "The river rose high on the bank.") A new generation of embeddings, however, produce vectors that are specific to the context in which the word is being applied, so-called "contextual word embeddings." For example, fixed embeddings assign the same vector to "play" occurring in "They played soccer" and occurring close to "They went to the play." With contextual word embeddings, once they are learned ("pre-trained") on giga-byte-scale dataset, they can assign a different embedding to each instance of "play" which better captures its sense based on the context. Unlike fixed word embeddings therefore, contextual word embeddings require context to be considered during extraction time (and not just during learning), and thus are computationally more intensive. While smaller scale versions of contextual embeddings have existed for decades (e.g. Leacock et al., 1993; Schwartz and Gomez, 2008; Dhillon et al., 2011), the recent wave of contextual embeddings are based on highly complex deep learning models such as bidirectional multi-layer recurrent neural networks (ELMO; Peters et al., 2018) or 12+ layer transformer networks (BERT: Kenton & Toutanova, 2019; XLnet: Yang et al., 2019; and RoBerta: Liu et al., 2019), which have led to dramatic improvement in performance in nearly all tasks they have been used including named entity recognition, question answering, automatic reading comprehension, dialog systems, machine translation, and sentiment analysis (Kenton & Toutanova, 2019; Peters, Neumann, Zettlemoyer & Yih, 2018). As of 2020, contextual embeddings have not been prominently used in the psychological literature.

Differential Language Analysis

LSA, LDA and the various embedding methods "cluster" language into lower dimensional representations of features. Differential Language Analysis, on the other hand, is a relatively simple method that explores the associations of language features with extra-linguistic author or text attributes of interest, such as personality. As such, it can use language clusters as features, or individual words and multi-word phrases. It is particularly useful for gaining insights into the words that best represent a construct. For example, relative frequencies for a given word can be derived and correlated with extraversion scores, resulting in a single correlation coefficient per word. The words and phrases that are most positively and negatively correlated with the outcome can then be shortlisted and visualized, yielding the language profile that most *differentiates* an outcome. As an open-vocabulary method, DLA is sensitive towards emoticons (*:-)*, ^_^), emojis and punctuations (*!!!!*), and misspellings, which is important for use with social media.[5] It also includes multi-word expressions (n-grams or phrases), or a set of words that

[5] Some closed-vocabulary dictionaries, such as LIWC2015, do include emoticons, common misspellings, and netspeak, but are limited by being static in nature and reflecting those that the developers were aware of. DLA better captures dynamic changes and idiosyncrasies of online language use.

commonly occur together (e.g., "happy new year"). (For a full overview of the method, see Schwartz et al., 2013b. For examples of DLA applied to personality, age, and gender, see Kern et al., 2014a; Kern et al., 2014b, and Park et al., 2016, respectively.)

Given its descriptive nature, this method works best on large datasets (we further consider and specify sample sizes below). DLA runs a large number of correlations. For instance, if a set of 1-to-3-grams has 20,000 words and phrases, 20,000 correlations are run. While the associated *p*-values are adjusted for multiple comparisons and can be used heuristically to identify potentially meaningful correlations, it is important to note that DLA fundamentally intended to be an exploratory method.

The Need for a Quantitative Comparison

Existing studies and reviews have indicated that both closed and open-vocabulary approaches have been used in psychological research to develop and test theory. Closed-vocabulary approaches can rapidly transform the thousands of mostly rarely used words in a given text sample into 10-100 interpretable language variables that can be explored with standard statistical techniques. As the derived language variables come from the same set of dictionaries, they are comparable across studies. However, closed vocabulary dictionaries are rigidly defined and insensitive to context and word sense. They are also unable to accommodate changing word senses over time. For example, LIWC2007 includes the word *sick* in the ***negative emotion*** and ***biological*** dictionaries. For many young people on social media in 2020, *sick* is a slang term that indicates that something is, in fact, fairly awesome. Such ambiguities can cause spurious correlations with dictionaries that are handled better by the open-vocabulary approaches.

Open-vocabulary approaches allow language variables to emerge from the data and are thus seemingly better suited for the discovery of language markers of novel psychological processes. From the possible clustering methods discussed above, we chose LDA topics for comparison as they are designed to be interpretable and semantically coherent as units of analysis, differentiate word senses and can provide nuance while still being relatively parsimonious. [6] However, open-vocabulary methods require more technical expertise in their implementation, require larger datasets, and are less convenient to use than the closed-vocabulary programs. As there are strengths and weakness of both approaches, it is important to consider the extent to which each approach is useful, under what conditions, and for what purposes.

Existing Comparisons

Correctly evaluating language analysis approaches is difficult. Both self-report questionnaires and language analyses seek to capture underlying, unobservable psychological characteristics, but neither adequately captures the "true" construct. To be useful for psychological research, language needs to be anchored to characteristics, with validity directly tested (e.g., Sun et al., 2019). The standard approach used to date is to treat self-reported data as the "ground truth," identifying the linguistic features that correlate with and/or predict different characteristics.

Using this approach, a number of reviews affirm the value of both closed- and open-vocabulary methods. Most previous reviews on automatic text analysis within psychology have focused on the various versions of LIWC. Tausczik and Pennebaker (2010) summarized the

[6] While methods exist to extract clusters of semantically close words from embedding spaces, we wanted to limit the comparison of exploratory methods to the single clustering approach mostly widely used in psychology. We do, however, report comparative personality prediction performances for LDA, word2vec and BERT embeddings in the prediction section.

relationships between LIWC2001 and LIWC2007 and the psychosocial processes associated with them. These included the connection between attentional focus and status hierarchy to pronouns, and function words to cognitive mechanisms. Pennebaker, Mehl, and Niederhoffer (2003) considered the association of LIWC2001 dictionaries with demographic, Big Five personality, and mental and physical health variables. Mehl (2006) summarized the different dictionary-based programs that preceded LIWC2001, including GI, DICTION, and TAS/C, providing a valuable introduction to closed-vocabulary approaches and emphasizing the power of the word count approach.

Despite the usefulness of the closed-vocabulary methods, Mehl's (2006) review also anticipated the power of more complex, machine-learning-based approaches. Reviews focused on open-vocabulary methods (e.g., Boyd & Pennebaker, 2015; Iliev et al., 2015; Schwartz & Ungar, 2015) suggest that text analysis methods range on a continuum from simple to complex– from human coders, to curated and crowd-sourced dictionaries, to the algorithmically derived language variables typical of open-vocabulary approaches. The reviews emphasize the potential of open-vocabulary approaches to lead to novel and unexpected advances based on "accidental discoveries" and underscore their enhanced predictive power.

Combining closed- and open-vocabulary approaches, Yarkoni's (2010) analysis of 694 bloggers tested associations between LIWC and word associations of lower-order personality facets, finding a variety of meaningful patterns. Schwartz et al. (2013b) tested machine-learning-based text prediction accuracies of personality for 75,000 Facebook users in the MyPersonality dataset, finding that language can moderately predict individual differences. Azucar, Marengo, and Settanni (2018) meta-analyzed prediction accuracies of Big Five traits from both text and other features, finding that predictive power was on par with standard behavioral predictors of personality.

The New Frontier: Online Text-Based Data

The largest modern sources of text are provided by social media, which capture a large fraction of users' behavior on the web (Gandomi & Haider, 2015; Kosinski et al., 2015). The rise of social media and other online data offers a new way of thinking for the social sciences. Over the past decade, many people have recorded their everyday thoughts, emotions, and behaviors in real-time. Unlike a questionnaire or lab-based study in which, for example, one's personality is measured and then correlated with a series of other measures, the online records allow consideration of how different characteristics are revealed across long time periods and a full range of contexts. Analysis of such text data is already playing a large role in psychological research (see Figure 3).

The claims and implications of these studies for psychological research and application depend on the extent to which they adequately capture psychological processes. To empirically inform best practices and clarify theoretical implications of different approaches, here we use the standard practice of assuming self-report as the ground truth and directly compare the results of the different open and closed-vocabulary approaches side-by-side.[7] To do this, we used the social media dataset that has been most widely used in psychological research: MyPersonality (Kosinski, Stillwell, & Graepel, 2013).

[7] Note that our goal here is to provide a comprehensive, empirical comparison of primary closed- and open-vocabulary approaches, describing our approach and providing codes, allowing replication to occur. For readers who are new to these methods, please see Kern et al., 2016 for specific guidance on extracting features, building models, and analyzing results.

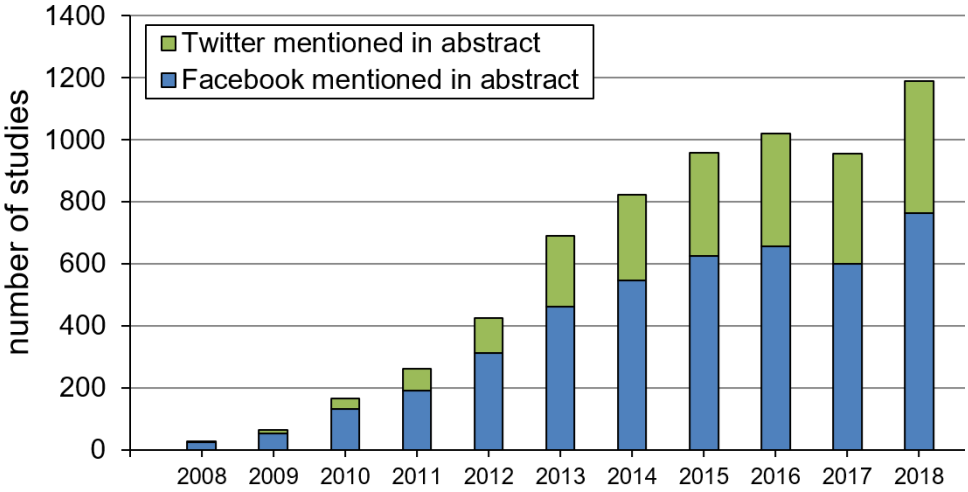


Figure 3. The number of studies indexed by PsycINFO mentioning Facebook (blue) or Twitter (green) in the abstract from 2008 to 2018 (as of April 2020).

Methods

The MyPersonality Dataset

MyPersonality was a third-party application on Facebook installed by roughly 4.5 million consenting users between 2007 and 2012 (Kosinski & Stillwell, 2012). The application allowed users to complete psychological inventories and to optionally share their results with friends. At a minimum, users completed 20 items from the International Personality Item Pool (IPIP; Goldberg et al., 2006), which assessed personality based on Costa and McCrae's (1992) five-factor model (the Big Five: extraversion, agreeableness, conscientiousness, neuroticism, openness to experience). All users agreed to the anonymous use of their survey responses for research purposes. A subset of the users also allowed the application to access their Facebook status messages. Age and gender, as reported within users' Facebook profiles, were also recorded, but comments on other users' statuses and updates shared by friends on their profiles were excluded from data collection.

A number of studies have used the dataset to predict Big Five personality from various "digital traces" (e.g., language, likes, or other online social interactions; see Azucar et al., 2018, for a meta-analysis of 12 such studies). Here, we compared the different closed- and open-vocabulary approaches in terms of their language correlates of gender, age, and personality, as well as their capacity to quantitatively capture variance in these traits. Our analysis implicitly assumes that gender, age, personality, and their manifestations in language are relatively stable over time, as the self-reported data were collected at a single time point, whereas language data stretched across several years.

We limited the sample to 65,896 individuals (62.07% female) who reported their age and gender, were between the ages of 16 and 60 years old (M=24.57 years, SD=9.01, median=21.00), completed the personality survey, and had at least 1,000 words across their status updates between January 2009 and November 2011. This amounted to over 12 million messages. Users wrote an average of 4,104 words across all status messages (median=2,875, SD=3,894, range=1,000 to 82,538).

Linguistic Feature Extraction

We transformed each user's collection of status messages into numerical variables that

captured the relative frequencies of three sets of language features: (a) words and phrases, (b) dictionaries, and (c) LDA topics.

Words and phrases. We first split users' statuses into **tokens**: single words including non-conventional usages and spellings (*omg*, *wtf*), punctuation, and emoticons (*:-]*, ^.^), using a social-media-appropriate tokenizer (Potts, 2011). We divided the frequencies of use for all tokens by each user's total number of tokens, yielding the users' relative frequencies of use.

Phrases – sequences of two (2-grams) and three (3-grams) tokens – capture distinctive language expressions that would otherwise be lost with single tokens (e.g., *happy birthday*, rather than *happy* and *birthday* or *sick of*, rather than *sick* and *of*). Rather than consider all possible combinations of two or three words that appear in a corpus, we considered only phrases that occurred with higher probability than the independent probabilities of their constituent words would suggest. For example, the phrase *happy birthday* was much more likely than the independent probabilities of *happy* and *birthday* would suggest. We used the pointwise mutual information (PMI) criterion to quantify these probabilities, keeping phrases with a threshold above 3 (for a full discussion, see Kern et al., 2016 and Schwartz et al. 2013b). Phrase frequencies were divided by the user's total number of words, yielding relative frequencies of each phrase.

As social media data include many idiosyncratic misspellings, plays on words, and borrowings from other languages, the vocabulary tends be larger than most other written texts; it is thus common to restrict analyses to words used by at least a certain fraction of the sample (e.g., Atkins et al., 2012). Accordingly, in DLA, we limited the analysis to tokens that were used by at least 5% of the users. This reduced the total number of distinct tokens from 1,680,708 to 2,986 words and 11,894 phrases.

Dictionaries. Once word frequencies have been extracted for a given user, the words can be matched against existing dictionaries to yield relative dictionary frequencies. Dictionary frequencies can be extracted using the programs themselves (DICTION, LIWC) or through a modern, Python-based codebase and MySQL infrastructure (DLATK, Schwartz et al., 2017; http://dlatk.wwbp.org). The former allows the previously-developed dictionaries to be used without modification, whereas the latter is easier to automate and can incorporate various improvements in the tokenization and handling of special language characters (e.g., emoticons, emojis). We used the simpler, program-based extraction method for our correlation analyses, both methods for the prediction analyses, and the DLATK dictionary extraction for our supplementary analyses.

We used the LIWC2015 software to extract the relative frequency of 73 primary LIWC dictionaries and 20 summary language variables for every user. DICTION was used to extract 31 DICTION dictionary frequencies, five master variables, and nine language statistics (see Supplementary Materials).[8] We used DLATK to extract the 182 GI dictionaries,[9] 31 DICTION dictionaries, 73 LIWC2015 dictionaries, and 64 LIWC2007 dictionaries (for supplementary analyses). We included multiple word endings as dictated by the dictionaries (e.g., *happ** included *happy*, *happier*, and *happiness*).

[8] We exported all the Facebook statuses and ran them through DICTION's batch mode in combinations of about 3,000 users at a time.

[9] Although GI's original 1960s implementations included rule-based routines to disambiguate words and account for word order, we only extracted the frequencies of GI dictionaries overall, as we believe that future users are more likely to use the dictionaries in a general-purpose word-counting software implementation.

Topic extraction. For DLA, we used a previously developed set of 2,000 Facebook topics, applying the existing topics to the current dataset. The topics were originally modeled using 14 million Facebook statuses (Schwartz et al., 2013b), and have been applied in subsequent studies with Facebook (e.g., Kern et al., 2014; Kern et al., 2014b; Park et al., 2015) and Twitter (Eichstaedt et al., 2015; Schwartz et al, 2013a). data (The topics can be downloaded at http://wwbp.org/data.)

We extracted the 2,000 topics from the language of every user in our dataset and multiplied the word-topic weights (*p(topic|word))*, which were determined during the modelling process with the relative frequencies of a users' words (*p(word|user)*), yielding the user's overall use of the topic:

$$p(topic|user) = \sum_{words \in topcis} p(topic|word) * p(word|user) \quad \text{eq. 2}$$

Each user received 2,000 topic scores, which we correlated with age, gender, and personality.

Analytic Approach

Our primary analyses involved correlational analyses across dictionaries, words, phrases, and topics, using the closed- and open-vocabulary approaches, with visualizations used to summarize results. Regression analyses compared predictive validity. We also considered necessary samples sizes and the utility of extracting different numbers of topics.

Correlational analyses. We used the 11,894 words and phrases, dictionaries, and the 2,000 topics as the dependent variables in separate regressions, with age, gender, and personality as predictors. Gender was controlled in age regressions; age was controlled in gender regressions, and both age and gender were controlled in personality regressions, with one personality factor tested at a time.

We used *p* values as a heuristic for identifying potentially meaningful correlations, acknowledging that analyses were exploratory and could be due to chance. Given the large number of regressions, we corrected for multiple comparisons using the Benjamini-Hochberg procedure (BH; Benjamini & Hochberg, 1995), which corrects the customary significance threshold (p=.05) for the number of features that are simultaneously being correlated. The BH procedure is less conservative but more powerful than corrections of the family-wise error rate, such as the Bonferroni correction (Holm, 1979), balancing between over- and under-estimating potential effects.

Visualizations. Word clouds are a space-efficient, information-dense way to visualize the most highly correlated words and phrases. In typical word clouds (e.g., www.wordle.net), the size of the word indicates the frequency of occurrence, and color is meaningless. We used DLATK to generate modified word clouds that scale the words by the magnitude of their correlation coefficient, such that larger words indicate stronger correlations with the outcome, and color indicates frequency, from red (frequently used) to blue (moderately used) to grey (rarely used). Thus, these modified word clouds summarize the words and phrases that most discriminate a given outcome while still providing an indication of frequency. To reduce repetition, we pruned duplicate mentions of a word (i.e., when a 1-gram also occurred in a phrase), giving preference to more highly correlated phrases over single words (cf. Schwartz et al., 2013c).

For topics, we created another type of modified word cloud, which shows the 10 words with the largest prevalence in the topic, with the size and color of the words scaled by descending prevalence (i.e., the largest, darkest word has the highest prevalence in the topic). Depending on the number of topics extracted, the LDA algorithm can create topics that are very similar to one another. To reduce repetition, we excluded topics from visualization if they shared

more than 25% of their top 15 words with the top 15 words of a more strongly correlated topic. Here we show the eight topics with the strongest associations after these exclusions.

Prediction. To quantify the amount of variance captured by the dictionaries and topics, we separately used each set of dictionaries and the 2,000 topics as features predicting gender, age, and personality. In choosing the prediction models, our goal was not necessarily to reach state of the art prediction performances (cf. Park et al., 2014; Sap et al., 2014; Schwartz et al., 2013b), but rather to use a predictive model that would be appropriate for both a relatively small (31 DICTION dictionaries) and large (2,000 LDA topics) number of features. We used penalized logistic regression (Gilbert, 2012) for the binary gender variable and penalized regression (or ridge regression; Hoerl & Kennard, 1970) for the continuous age and personality variables. Both techniques are straight-forward machine learning extensions of logistic and linear regression, where the squared magnitude of the coefficients is added as a penalty to the error function, which addresses problems of collinearity between the coefficients (language features are often highly intercorrelated) and reduces overfitting the model to the specific dataset (Fan et al., 2008).

To determine prediction accuracies, we used 10-fold validation. The data are randomly split into ten subsets ("folds"), and a model is fit over nine of the folds ("training set"). The trained model is then applied to the remaining fold ("test set"), and its predicted outcome values (e.g., user extraversion scores) are compared to the actual, user-reported values. Accuracy is calculated as the Pearson correlation between the predicted and actual outcome values. This procedure is then repeated in round-robin fashion until every fold serves as the test set once. The final predictive accuracy is the average of the 10 test set accuracies.

Power analyses: Sample size and words per user. One advantage of closed-vocabulary methods is their relatively small number of language features (i.e., a limited set of dictionaries), which can increase their power in exploratory analyses may be more parsimonious than the large number of features in the open-vocabulary methods. To inform which method is appropriate for datasets of different sizes, we repeated the exploratory language analyses across randomly-selected samples of 50, 500, 1,000, 2,000, 5,000, 15,000, and 50,000 users. Separately, we also explored how many words are needed from a given user to produce nuanced profiles of language associations. The average Facebook status had a length of 21.45 1-grams in our data set, and so we sampled the most recent 1, 2, 4, 7, 10 statuses from users, yielding the most recent 21, 43, 86, 150, 214, 300, 515, 751, and 1,008 words across random samples of N = 150, 1,000 and 5,000 users.

Choosing the number of topics to extract. In the LDA topic modeling process, the numbers of topics to extract (*k*) needs to be specified. To inform what *k* is optimal, we used LDA to model 50, 500, and 2,000 topics across random subsets of the Facebook dataset comprised of 50, 500, 5000, 50,000, 500,000, and five million statuses. This yielded a total of 18 sets of topics (three choices for number of topics * six status sizes). We first examined the ability of the 50, 500, and 2000 topics modeled over five million statuses to distinguish contexts and word-senses of the word *play*, a word commonly used in different contexts. Then, to quantify the information captured by the different number of topics, we used the 18 sets of extracted topic frequencies as features in 18 machine learning prediction models (using ridge-regression), predicting age, gender, and personality of the users, and report the average cross-validated prediction accuracies as a measure of how much nuance is captured by the different sets of topics.

Results

Comparing the Three Closed-Vocabulary Programs

The GI, DICTION, and LIWC dictionaries cover similar concepts, but also reflect the

different purposes for which they were developed. Despite differences in purpose, all three programs include positive affect, negative affect, and first-person singular pronoun dictionaries. As can be seen in Table 1, the frequencies of these dictionaries are significantly correlated with one another across programs, and with similar dictionaries within the same program. These inter-correlations are largely due to overlap in the words that the dictionaries contain. A few very frequent words often contribute the majority of counts in dictionaries (see Supplementary Material for the most frequent words in the dictionaries); when they occur in multiple dictionaries, these dictionaries will be highly correlated. Thus, it is not surprising that function word dictionaries with a few highly frequent words (e.g., *the*, *and*, *to*) have the strongest correlations across programs.

Other dictionary concepts that are covered across programs include cognition and complexity of language (Harvard-IV ***abstract vocabulary***; DICTION ***cognition***; LIWC ***insight***, ***tentative***, ***causation***, ***cognitive processes***; Lasswell ***enlightenment*** dictionaries,), as well as economic and fiscal concerns (Harvard-IV ***economic***; Lasswell ***wealth*** dictionaries; LIWC ***money***, ***work***, ***achievement***).

Table 1

Intercorrelations Amongst Positive Affect, Negative Affect, and Pronoun Dictionaries.

	General Inquirer			Diction		LIWC 2015
	Lasswell	Harvard IV	Osgood			
	Positive Affect	Pleasure	Positive	Optimism	Satisfaction	Affect
General Inquirer						
Pleasure	.48					
Positive	.70	.63				
Diction						
Optimism	.33	.45	.33			
Satisfaction	.31	.53	.34	.72		
LIWC						
Affect	.37	.47	.33	.27	.37	
Pos. Emotion	.45	.60	.42	.46	.45	.85

	General Inquirer				Diction		LIWC 2015	
	Lasswell	Harvard IV	Stanford					
	Negative Affect	Vice	Negative	Hostile	Hardship	Blame	Swear	Negative Emotion
General Inquirer								
Vice	.59							
Negative	.68	.76						
Hostile	.60	.54	.85					
Diction								
Hardship	.26	.23	.26	.17				
Blame	.27	.27	.22	.14	.12			
LIWC								
Swear	.39	.26	.38	.37	.13	.10		
Negative Emotion	.56	.45	.49	.34	.36	.28	.61	
Anger	.48	.37	.46	.41	.24	.17	.87	.76

	Gen. Inquirer	Diction	LIWC 2015	
	Harvard IV: Self	Self-reference	Pronouns	Pers. pronouns
Diction				
Self-reference	.75			
LIWC				
Pronouns	.75	.49		
Pers. Pronouns	.70	.60	.96	
1st. pers. sing.	.92	.80	.75	.77

Note. DICTION and LIWC2015 dictionaries were extracted through their respective programs, GI dictionaries through DLATK.

Language Profiles of Gender, Age, and Personality

Figure 4 provides a quantitative summary of the correlates of gender, age, and personality across the five methods. The figure provides the ten largest positive and negative standardized regression coefficients between the dictionaries and outcomes[10] and the most strongly associated topics, words, and phrases.

Gender. As summarized in Figure 4a, the GI ***female*** and LIWC ***female references*** dictionaries were strongly correlated with female gender. Identifying as female was associated with dictionaries capturing positive emotion, first-person pronouns, and language associated with close relationships. Similarly, in the DLA word clouds, female gender was correlated with high-arousal emotions (*excited*, *happy*, *yay!*) and mentions of *love*.

Identifying as male was associated with dictionaries reflecting negative emotion, economic concerns, and hostility and aggression. The GI-Stanford dictionaries clearly separate the genders along the ***affiliative-passive-positive*** (female) and ***hostile-strength-negative*** (male) dimensions. Male gender was also associated with the use of articles and prepositions in the LIWC dictionaries, as well as the most-associated open-vocabulary words (*of, the, in, by*). The LDA topics further reveal that male-associated words reflect economic concerns, such as *tax, budget, economy, government, income,* and *benefits*, and that male language associations with hostility and aggression may in large part be specifically driven by competition (*battle, victory, fight*), political debate (*country, power, freedom*), and sports (*football, season, team*; *win, lose, bet*).

Age. As summarized in Figure 4b, younger age was associated with self-reference and negative emotion. Older age was associated with mentions of others, economic concerns, and family and social categories. Similar themes appear in the LDA topics, with older age most strongly associated with friend and family topics. Older individuals also tended to use longer sentences and more function words, which was mirrored in the DLA dominate use of function words. The DLA word clouds mark younger age by the use of emoticons, colloquialisms, and contractions, and suggest *hate, bored,* and *stupid* as specific expressions of negative emotions.

Personality. Associations between personality and language variables (typically $|\beta|<.15$) were weaker than those for age and gender (typically $|\beta|<.30$). Across personality dimensions, the strongest associations were generally with positive and negative emotion dictionaries.

Agreeableness demonstrated the strongest associations with positive emotion. It was weakly associated with greater use of first-person plural pronouns, and with dictionaries reflecting affiliation. Low agreeableness was dominated by swear words. DLA across topics, words, and phrases reveal high agreeableness to be marked by expression of delight and gratitude (*wonderful, amazing, thank you*), social connection and events (*friends, family, weekend, thanksgiving*), and religiosity. The language of disagreeableness included cursing and negative appraisals of others (*rude*, *selfish*, *ignorant*).

Conscientiousness was positively associated with references to work and economic concerns, references to time, and social connection. DLA topics revealed that conscientious language includes references to family and friends (*family, friends, blessed*), structured social time (*weekend, spending, hanging*), and relaxing from work (*relaxation, vacation, recover*).

---

[10] When reporting dictionary correlations, we took into account the hierarchical structure of the dictionaries (e.g., words in the LIWC *anger* dictionary are part of the LIWC *negative emotion* dictionary). If the broader dictionary showed a significant association, we noted the sub-dictionaries as well. If the broader dictionary did not show a significant association but two or more sub-dictionaries were significant, we note the higher order dictionary but leave the coefficient blank.

Individuals low in conscientiousness were more likely to use curse words.

Extraversion was weakly associated with the emotion and social dictionaries. DLA emphasized social events. Low extraversion predominantly focused on computers and technology, Japanese culture (*anime, manga, episode*) and books and reading, which are concepts that are not well captured by any dictionary.

Neuroticism was most distinguished by its association with negative emotion dictionaries, and inversely with positive emotions. The most strongly associated DLA topics reflected somatic concerns (*feeling, tired, sick),* hostility and cursing, exhaustion and over-arousal (*stressed, frustrated, annoyed,)* and depressed mood. Emotional stability (low neuroticism) was distinguished by mentions of weekends (*awesome, weekend, amazing*), sports, and religion.

Openness was positively associated with cognitive dictionaries, reflecting intellect and insight, and syntactic markers of increased sentence complexity. DLA topic correlations reflected existential (*human, nature, universe, wonders*) and artistic (*writing, write, poetry*) concerns. Low openness was associated with pragmatic, domestic concerns including home, family, and temporal concepts.

LIWC2007 vs. LIWC2015. The LIWC2007 dictionaries have most often been used in psychological research, but have been replaced by the 2015 version; our comparisons are based upon this more updated version. As a supplemental analysis, we repeated the analyses using the 2007 dictionaries (see Supplemental Materials). Dictionaries covering the same concept or part of speech (e.g., pronouns) demonstrated very similar patterns of association. The 2015 dictionaries added several dictionaries that correlate with gender and personality, including ***female references***, ***Netspeak***, ***time orientation***, and different ***drive*** dictionaries.

Figure 4. Standardized regression coefficients between user age and dictionaries (top), topics (bottom left), and words and phrases (bottom right) across gender (3a), age (3b), and personality (3c-g) outcomes. Age associations are controlled for gender, gender for age, and personality for both.

A) Gender

	General Inquirer						DICTION		Linguistic Inquiry and Word Count (LIWC 2015)			
	Lasswell		Harvard IV		Stanford				LIWC (other)		LIWC (psych. processes)	
	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β
Female	Affect		Pleasure	.29	Affiliation	.12	Optimism (m)	.14	Emotional tone (m)	.27	Social processes	.12
	Affect-Other	.28	Females	.28	Passive	.09	+Satisfaction	.22	Personal pronoun	.17	Female reference	.30
	Affect-Domain	.21	Emotion	.25	Positive	.09	+Praise	.08	1st pers singular	.16	Family	.28
	Affect-Gain	.16	Kinship	.20	Weak	.06	+Inspiration	.05	3rd pers singular	.11	Affective process	.25
	Affect-Participants	.05	Self	.15	Submit	.05	-Blame	.04	2nd person	.07	Positive emotion	.29
	Wellbeing-Total	.15	Children	.15			Certainty (m)		Total pronouns	.11	Home	.21
	Wellbeing-Psych.	.24	Independent Adj.	.12			+Insistence	.07	Common adverbs	.09	Netspeak	.18
	Wellbeing-Participants	.16	State Verb	.12			-Self-reference	.15	Common verbs	.07	Affiliation	.17
	Positive-Affect	.11	Need	.11			+Tenacity	.06	Conjunctions	.07	Future focus	.10
	Transaction-Gain	.10	Evaluation 2	.10			Human Interest	.12	Common adjectives	.06	Nonfluencies	.10
	Respect-Lose	.07					Temporal	.05				
Male	Wealth-Total	.19	Military	.21	Strength	.09	Realism		Articles	.24	Death	.22
	Wealth-Other	.19	Movement-Exert	.21	Hostile	.08	+Familiarity	.09	Analytical thinking (m)	.19	Anger	.21
	Power-Total	.18	Political	.19	Negative	.07	+Spatial	.09	Comparisons	.12	Drives	
	Power-Arenas	.15	Economic	.16	Understated	.06	-Complexity	.08	Prepositions	.12	Power	.20
	Power-Conflict	.14	Region	.15	Active	.06	Activity		Impersonal pronouns	.08	Achievement	.13
	Power-Participants	.14	Space	.15	Power	.06	+Aggression	.10	Quantifiers	.06	Risk	.09
	(Ordinary)		Doctrine	.15			+Accomplishment	.07	Interrogatives	.06	Swear words	.19
	Power-Authority	.13	Abstract vocab.	.14			+Communication	.07	3rd pers plural	.05	Sexual	.19
	Power-Loss	.12	Collectives	.14			Commonality		Numbers	.04	Space	.16
	Arenas	.17	Expressive	.13			+Centrality	.08			Money	.11
	Religion	.14					-Diversity	.06			Tentative	.09
							-Exclusion	.05				
							Collectives	.06				

LDA Topics Most Associated with Female

β = .30 β = .28 β = .28 β = .28 β = .26 β = .25 β = .25 β = .24

LDA Topics Most Associated with Male

β = .24 β = .22 β = .22 β = .21 β = .21 β = .20 β = .19 β = .19

prevalence in topic

50 Words and Phrases Most Associated with Female

Male

Note. All coefficients are significant at $p < .001$, corrected for multiple comparisons. (m) designates "master" categories that combine frequencies of multiple dictionaries.

B) Age

	General Inquirer						DICTION		Linguistic Inquiry and Word Count (LIWC 2015)			
	Lasswell		Harvard IV		Stanford				LIWC (other)		LIWC (psych. processes)	
	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β
Older	Power-Total	.17	Kinship	.29	Power	.16	Realism	.16	Clout (m)	.29	Social processes	.19
	Power-Other	.23	Economic	.25	Positive	.12	+Familiarity	.24	Articles	.29	Family	.27
	Power-Participants	.17	Communication Tools	.24	Affiliation	.11	+Human Interest	.21	Prepositions	.28	Drives	
	(Authority)		Human	.21	Submit	.09	-Complexity	.11	Quantifiers	.24	Affiliation	.21
	Wealth-Total	.22	1st pers. plural	.20	Strength	.04	Certainty	.23	Emotional tone(m)	.21	Power	.20
	Wealth-Other	.19	Political	.18	Understated	.04	+Collectives	.11	Analytical thinking (m)	.21	Relativity	.14
	Transaction-Gain	.19	Region	.18	Overstated	.02	+Insistence	.10	Personal pronouns		Space	.21
	Respect-Other	.20	Role	.17			+Tenacity	.09	3rd pers plural	.24	Personal concerns	
	Means	.18	Objects	.17			Rapport	.16	1st pers plural	.18	Money	.20
	Affect-Participants	.18	Male	.16			Optimism	.12	3rd pers singular	.13	Religion	.18
	Wellbeing-Gain	.16							Function words	.13	Home	.17
Younger	Negative-Affect	.24	Self	.20	Negative	.19	Certainty		Personal pronouns	.14	Affective process	.20
	Affect-Gain	.18	Academic vocab.	.19	Hostile	.16	-Self-reference	.22	1st pers singular.	.27	Negative emotion	.33
	Wellbeing-Loss	.17	Emotion	.16	Passive		-Ambivalence	.03	Negations	.18	Anger	.27
	Rectitude-Gains	.12	Pain	.14			-Variety	.05	Common Adverbs	.17	Sadness	.17
	Enlightenm.-Ends	.12	Disagreement	.14			Optimism		Pronouns	.08	Informal language	
	Transaction-Loss	.09	Vice	.14			-Hardship	.12	Authentic(m)	.07	Netspeak	.30
	Power-Conflict	.08	Expressive	.12			-Blame	.11	Numbers	.05	Swear words	.21
	Affect-Loss	.07	Nature Process	.11			-Denial	.04			Assent	.15
	Enlightenm.-Other	.06	Say	.10			Present-Concern	.03			Nonfluencies	.14
	Denial	.06	Very	.09			Activity				Biological process	
							-Cognition	.03			Body	.17
							+Aggression	.03			Sexual	.16
							+Motion	.02				

LDA Topics Most Associated with Older Age

LDA Topics Most Associated with Younger Age

prevalence in topic

50 Words and Phrases Most Associated with Older Age

Younger Age

Note. All coefficients are significant at $p < .001$, corrected for multiple comparisons. (m) designates "master" categories that combine frequencies of multiple dictionaries.

C) Agreeableness

	General Inquirer						DICTION		Linguistic Inquiry and Word Count (LIWC 2015)			
	Lasswell		Harvard IV		Stanford				LIWC (other)		LIWC (psych. processes)	
	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β
High Agr.	Wellbeing-Psychological	.08	Pleasure	.12	Positive	.10	Optimism	.11	Emotional tone (m)	.21	Positive emotion	.14
	Affect-Other	.07	Time Broad	.08	Affiliation	.06	+Satisfaction	.08	Clout (m)	.07	Drives	.09
	Positive-Affect	.07	Religion	.07	Overstated	.05	+Praise	.07	Personal pronouns		Affiliation	.09
	Certainty	.06	1st pers. plural	.06	Power	.04	+Inspiration	.05	1st pers plural	.06	Reward	.06
	Space-Time	.06	Virtue	.06	Strength	.04	Certainty	.06	Common adjectives	.05	Achievement	.05
	Rectitude-Ends	.06	Expressive	.05	Submit	.03	+Leveling	.03	Prepositions	.04	Relativity	.07
	Transaction-Gain	.05	Quantity Ordinal	.04	Passive	.02	+Insistence	.06	Quantifiers	.04	Time	.08
	Respect-Total	.05	Names	.04	Understated	.02	Realism	.04	Authentic (m)	.03	Motion	.05
	Respect-Lose	.04	Independent Adj.	.04			+Temporal	.05			Future focus	.07
	Skill-Aesthetic	.04	Sky	.04							Religion	.07
Low Agr.	Negative-Affect	.09	Vice	.08	Negative	.08	Optimism		Negations	.06	Negative emotion	.15
	Wellbeing-Loss	.05	Disagreement	.06	Hostile	.06	-Hardship	.05	Personal pronouns		Anger	.20
	Denial	.04	Negation	.04			-Blame	.04	1st pers singular	.03	Anxiety	.03
	Power		Races	.03			-Denial	.03	3rd pers plural	.03	Swear words	.16
	Power-Authority	.03	Increase	.03			Aggression	.04			Biological process	.05
	Power-Participants (Authority)	.03	Say	.03			Variety	.03			Sexual	.13
			Color	.03			Self-reference	.02			Body	.08
	Power-Participants (Ordinary)	.03	Nature Process	.03			Communication	.02			Personal concerns	
			Body Parts	.03							Death	.10
	Negative-Value	.03	Movement-Exert	.03							Money	.04
	Affect-Loss	.03									Risk	.05
	Wealth-Transaction	.03										
	Skill-Participant	.02										

LDA Topics Most Associated with High Agreeableness

β = .12 β = .12 β = .11 β = .10 β = .10 β = .09 β = .09 β = .09

LDA Topics Most Associated with Low Agreeableness

β = .14 β = .09 β = .09 β = .09 β = .08 β = .07 β = .70 β = .07

b b b
prevalence in topic

50 Words and Phrases Most Associated with High Agreeableness

a a a
correlation strength

Low Agreeableness

relative frequency

Note. All coefficients are significant at $p < .001$, corrected for multiple comparisons. (m) designates "master" categories that combine frequencies of multiple dictionaries.

D) Conscientiousness

	General Inquirer						DICTION		Linguistic Inquiry and Word Count (LIWC 2015)			
	Lasswell		Harvard IV		Stanford				LIWC (other)		LIWC (psych. processes)	
	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β
High Con.	Affect-Other	.08	Time Broad	.10	Positive	.09	Certainty	.07	Emotional tone (m)	.17	Drives	.12
	Space-Time	.07	Pleasure	.09	Strength	.07	+Insistence	.09	Prepositions	.07	Achievement	.12
	Transaction-Gain	.07	Economic	.07	Affiliation	.07	+Collectives	.05	Clout (m)	.06	Reward	.09
	Certainty	.07	Strength	.07	Power	.05	Optimism	.08	Quantifiers	.06	Affiliation	.07
	Wellbeing-Psych.	.06	Travel	.07	Submit	.05	+Satisfaction	.05	Personal pronouns		Relativity	.10
	Skill-Other	.06	Virtue	.07	Overstated	.04	+Praise	.05	1st pers plural	.05	Time	.11
	Ends	.06	Comparison	.06	Active	.04	Realism	.07	Analytical thinking(m)	.05	Motion	.06
	Nations	.05	1st Person Plural	.06	Understated	.02	+Temporal	.07	Common adjectives	.05	Positive emotion	.11
	Skill-Total	.05	Interpretative	.06			+Familiarity	.05	Authentic (m)	.04	Work	.11
	Positive-Affect	.05	Action	.06			Accomplishment	.06	Articles	.03	Future focus	.07
Low Con.	Negative-Affect	.06	Vice	.07	Negative	.05	Certainty		Personal pronouns	.04	Negative emotion	.12
	Wellbeing-Loss	.05	Color	.05			-Self-reference	.05	1st pers singular	.06	Anger	.13
	Enlightenm-Ends	.05	Self	.04			-Variety	.04	Negations	.05	Sadness	.05
	Affect-Loss	.03	Disagreement	.04			-Ambivalence	.02			Biological process	.06
	Negative-Value	.02	Think	.04			Optimism				Sexual	.11
	Form	.02	Tool	.03			-Hardship	.04			Body	.09
			Races	.03			-Blame	.03			Swear words	.10
			Evaluation2	.03			-Denial	.02			Death	.09
			Nature Process	.03			Communication	.03			Perceptual process	.08
							Exclusion	.02			Hear	.07
							Past-Concern	.01				

LDA Topics Most Associated with High Conscientiousness

β = .13 β = .13 β = .12 β = .12 β = .12 β = .12 β = .12 β = .12

LDA Topics Least Associated with Low Conscientiousness

β = .10 β = .08 β = .08 β = .07 β = .07 β = .06 β = .06 β = .07

prevalence in topic

50 Words and Phrases Most Associated

High Conscientiousness

Low Conscientiousness

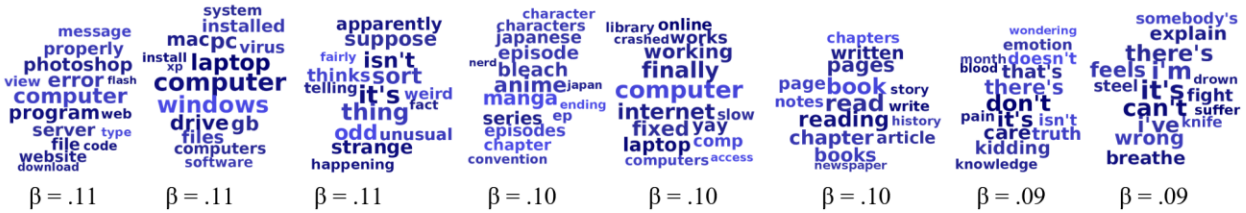
Note. All coefficients are significant at $p < .001$, corrected for multiple comparisons. (m) designates "master" categories that combine frequencies of multiple dictionaries.

E) Extraversion

	General Inquirer						DICTION		Linguistic Inquiry and Word Count (LIWC 2015)			
	Lasswell		Harvard IV		Stanford				LIWC (other)		LIWC (psych. processes)	
	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β
High Ext.	Affect-Total		Pleasure	.12	Affiliation	.09	Optimism	.11	Emotional tone (m)	.18	Positive emotion	.16
	Affect-Other	.14	Children	.07	Positive	.08	+Satisfaction	.08	Clout (m)	.06	Drives	
	Affect-Domain	.12	Vary	.06	Strength	.04	+Praise	.05	Personal pronoun	.03	Affiliation	.12
	Affect-Gain	.10	Movement-Rise	.06	Active	.02	+Inspiration	.03	2nd person	.04	Reward	.09
	Affect-Participants	.04	Completion	.06			Insistence	.06	1st pers plural	.03	Netspeak	.11
	Positive-Affect	.07	Names	.06			Realism	.01	1st pers singular	.02	Social processes	.05
	Nations	.05	Emotion	.05			+Human Interest	.02			Friends	.09
	Power-Participants	.04	Travel	.05			+Temporal	.02			Family	.05
	(Ordinary)		Social Relation	.05			+Spatial	.02			Leisure	.07
	Wellbeing-Psych.	.04	Movement-Change	.05			Self-reference	.01			Future focus	.05
	Power-Cooperation	.04									Biological processes	.04
	Transaction-Gain	.04										
Low Ext.	Enlightenm.-Total	.06	Negation	.09	Weak	.05	Denial	.06	Negations	.06	Personal concern	
	Enlightenm.-Other	.08	Awareness	.08	Negative	.03	Hardship	.06	Auxiliary verbs	.06	Death	.10
	Enlightenm.-Ends	.08	Vice	.07	Understated	.03	Tenacity	.06	Personal pronouns		Work	.05
	Enlightenm.-Part.	.05	Abstract vocab.	.06			Ambivalence	.05	3rd pers plural	.06	Cognitive process	.09
	Denial	.06	Doctrine	.96			Activity		Impersonal pronouns	.05	Tentative	.09
	Uncertainty	.05	Comm. Tools	.06			-Cognition	.05	Common verbs	.05	Insight	.09
	Affect-Loss	.05	Change Finish	.05			+Communication	.03	Common adverbs	.05	Differentiation	.08
	Means	.05	Academic vocab.	.05			+Aggression	.03	Articles	.04	Causation	.07
	Negative-Value	.04	Pain	.05			Complexity	.04	Comparisons	.04	Risk	.08
	Negative-Affect	.04	Cardinal	.05			Familiarity	.04	Interrogatives	.04	Negative emotion	.07
							Exclusion	.04			Anxiety	.07

LDA Topics Most Associated with High Extraversion

LDA Topics Most Associated with Low Extraversion



50 Words and Phrases Most Associated with High Extraversion

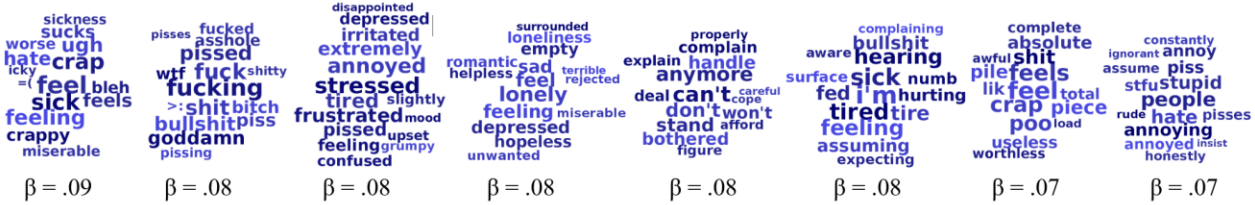
Low Extraversion

Note. All coefficients are significant at $p < .001$, corrected for multiple comparisons. (m) designates "master" categories that combine frequencies of multiple dictionaries.

F) Neuroticism

	General Inquirer						DICTION		Linguistic Inquiry and Word Count (LIWC 2015)			
	Lasswell		Harvard IV		Stanford				LIWC (other)		LIWC (psych. processes)	
	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β
High Neu.	Negative-Affect	.10	Pain	.09	Weak	.07	Optimism		Negations	.07	Negative emotion	.15
	Affect-Loss	.06	Vice	.09	Negative	.07	-Hardship	.07	Common adverbs	.05	Anger	.11
	Wellbeing-Total	.03	Weak	.07	Passive	.04	-Blame	.05	Common verbs	.05	Sadness	.09
	Wellbeing-Loss	.05	Negation	.06	Hostile	.03	-Denial	.04	Personal pronouns	.03	Anxiety	.08
	Wellbeing-Phys.	.03	Need	.04	Understated	.02	Certainty		1st pers singular	.05	Death	.08
	Denial	.05	Self	.04			-Ambivalence	.05	3rd pers singular	.02	Cognitive process	.06
	Enlightenm.-Ends	.05	State Verb	.04			-Self-reference	.04	Auxiliary verbs	.04	Discrepancy	.07
	Negative-Value	.04	Awareness	.04			+Tenacity	.03	Conjunctions	.03	Tentative	.06
	Rectitude-Ethics	.03	Change-Finish	.04			Exclusion	.03			Biological processes	
	Enlightenm.-Other	.03	Disagreement	.03			Aggression	.02			Body	.06
							Present-Concern	.02			Sexual	.06
							Communication	.02				
Low Neu.	Affect-Other	.05	Pleasure	.07	Positive	.06	Optimism	.09	Emotional tone (m)	.17	Positive emotion	.10
	Nations	.04	Ritual	.07	Affiliation	.04	+Praise	.04	Clout (m)	.06	Drives	
	Power		Expressive	.06	Strength	.03	+Satisfaction	.03	Analytical thinking (m)	.06	Affiliation	.07
	Power-Coop.	.04	Places	.05	Power	.02	+Inspiration	.03	Personal pronouns		Reward	.07
	Power-Part. (Ordinary)	.04	1st pers. plural	.05			Certainty	.03	1st pers plural	.05	Achievement	.06
			Names	.04			+Insistence	.05	Articles	.03	Personal concern	
	Power-Conflict	.03	Political	.04			+Collectives	.03			Leisure	.07
	Positive-Affect	.04	Land Places	.04			Temporal	.03			Religion	.05
	Respect-Lose	.03	Time-Broad	.04			Spatial	.02			Netspeak	.05
	Rectitude-Ends	.03	Travel	.04			Cooperation	.02			Relativity	.04
	Affect-Domain	.03									Time	.04
	Skill-Total	.03									Motion	.04

LDA Topics Most Associated with High Neuroticism



LDA Topics Most Associated with Low Neuroticism

50 Words and Phrases Most Associated with High Neuroticism

Low Neuroticism

Note. All coefficients are significant at $p < .001$, corrected for multiple comparisons. (m) designates "master" categories that combine frequencies of multiple dictionaries.

G) Openness

	General Inquirer						DICTION		Linguistic Inquiry and Word Count (LIWC 2015)			
	Lasswell		Harvard IV		Stanford				LIWC (other)		LIWC (psych. processes)	
	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β
High Ope.	Skill-Aesthetic	.10	Awareness	.12	Understated	.06	Certainty		Articles	.15	Cognitive process	.09
	Enlightenm-Total	.07	Abstract vocab.	.10	Negative	.04	-Variety	.09	Total function words	.08	Insight	.12
	Enlightenm-Other	.09	Think	.09	Overstated	.04	+Tenacity	.07	Auxiliary verbs	.07	Causation	.07
	Enlightenm-Ends	.06	Doctrine	.08	Weak	.03	-Self-reference	.04	Comparisons	.06	Tentative	.07
	Arenas	.08	Quality	.08	Passive	.02	-Ambivalence	.04	Impersonal pronouns	.06	Death	.12
	Form	.07	Perceive	.07			Complexity	.11	Conjunctions	.06	Perceptual process	.12
	Power-Authority	.06	Nature-Process	.07			Familiarity	.07	Prepositions	.05	Hear	.08
	Power-Participants	.05	Independent Adj.	.07			Cognition	.06	1st pers singular	.04	See	.07
	(Ordinary)		Negation	.07			Centrality	.06	Interrogatives	.04	Anxiety	.08
	Wealth-Total	.05	Evaluation2	.07			Exclusion	.06	Quantifiers	.04	Space	.05
	Wealth-Other	.06					Communication	.04				
Low Ope.	Affect		Kinship	.10	Submit	.07	Certainty	.02	Emotional tone (m)	.08	Netspeak	.14
	Affect-Other	.08	Persistence	.10	Affiliation	.04	+Insistence	.13	Clout (m)	.05	Family	.13
	Affect-Participants	.05	Pleasure	.08			+Collectives	.02	2nd person	.02	Affective process	.10
	Affect-Domain	.05	Time (Broad)	.07			Realism	.02			Positive emotion	.11
	Power-Cooperation	.07	Movement-	.07			+Temporal	.07			Drives	
	Well-being Total	.04	Change (Stay)				+Human Interest	.02			Reward	.11
	Wellbeing-Psych.	.07	Social	.06			Optimism	.04			Affiliation	.08
	Wellbeing-Participants	.04	Ritual	.06			+Satisfaction	.04			Future focus	.09
	Respect-Lose	.06	Try	.05			+Praise	.03			Home	.08
	Positive-Affect	.05	Vary	.05			Motion	.04			Relativity	.05
	Nations	.04	Travel	.05							Time	.10

LDA Topics Most Associated with High Openness

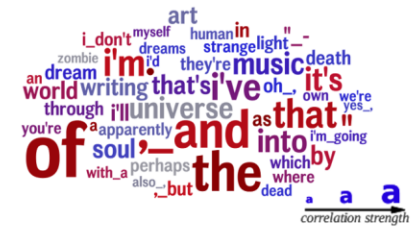
β = .18 β = .17 β = .16 β = .13 β = .13 β = .13 β = .12 β = .12

LDA Topics Most Associated with Low Openness

β = .11 β = .11 β = .11 β = .11 β = .11 β = .11 β = .10 β = .10

prevalence in topic

50 Words and Phrases Most Associated with High Openness



Low Openness

relative frequency

Note. All coefficients are significant at $p < .001$, corrected for multiple comparisons. (m) designates "master" categories that combine frequencies of multiple dictionaries.

Predictive Power

To quantitatively gauge how much each approach captures variance in gender, age, and personality, we examined the cross-validated prediction performances of models that used the different sets of language variables as features and compared them with the accuracies of previously published prediction models that combined topics, words, and phrases as features on the study dataset (Park et al., 2014; Sap et al., 2014). For comparison with more recent methods, we reported prediction accuracies based on Word2Vec word embeddings and contextual BERT embeddings also obtained on the study dataset (Lynn, Balasubramanian & Schwartz, 2020). Finally, we include Azucar et al.'s (2018) meta-analytic estimates for prediction accuracies for social media-based prediction of Big Five personality across data sets.

As shown in Table 2, DICTION's dictionaries captured less information about personality ($r_{average}$= .23) than the LIWC ($r_{average}$= .28) and GI dictionaries ($r_{average}$= .29). As LIWC includes about a third of the dictionary categories of GI, it appears more parsimonious while equally exhaustive.

The LDA topic predictions were about 30% higher than those achieved by GI and LIWC and almost indistinguishable from more sophisticated prediction models using many more language features (including words and phrases). The adjusted R^2 for LIWC, GI, and the LDA topics was comparable (R^2=.08, .08, .11, respectively). The average personality prediction accuracies for the models based on 2,000 topics with and without additional features, Word2Vec and BERT embeddings were very similar ($r_{average}$= .37 to 39) and nominally above the meta-analytic baseline ($r_{average}$= .35). This suggests that all these approaches capture a similar amount of language variance -- but that particularly the word embeddings do so more parsimoniously with fewer language dimensions (200).

Impact of Sample Size

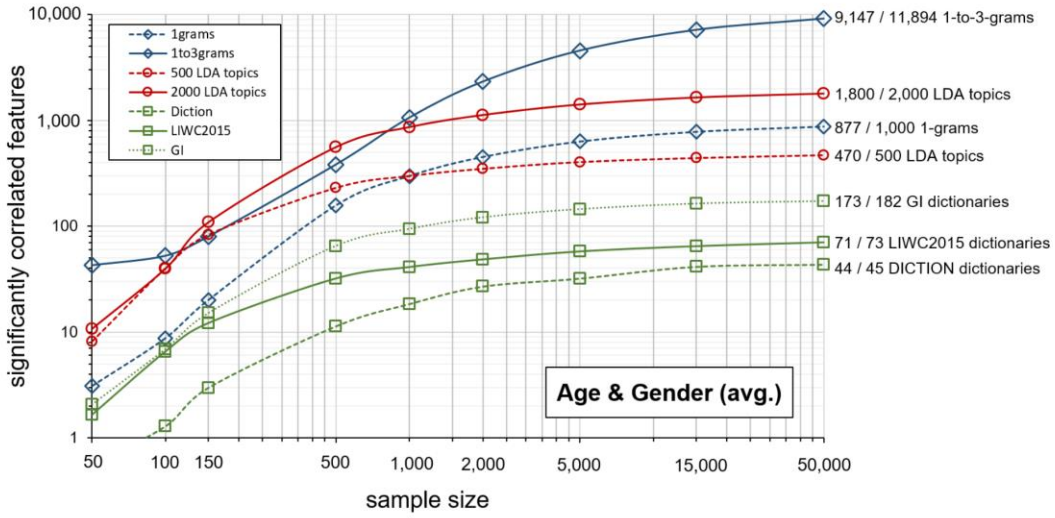
Figure 5 shows how many language features are significantly associated (after BH correction) with age and gender (combined), and personality (averaged across the five traits) as a function of different sample sizes (see Supplementary Material for each outcome). As a rough guide, theoretically interesting findings occurred with about 10 LIWC dictionaries, 100 LDA topics or 200 words and phrases. As shown in Table 3, while a few hundred users were sufficient for age and gender, much larger samples were needed for personality. There was variance between the traits; for example, for openness, 550 users sufficed for 100 significantly associated LDA topics, whereas for neuroticism, a sample of 1,800 was needed.

Table 2

Cross-validated Prediction Performances of Prediction Models Using the Dictionaries of the Different Software Programs.

	Diction	LIWC 2015	Gen. Inquirer	LDA Topics	LDA Topics, Words, Phrases	Word2Vec Embeddings	BERT Embeddings	Meta-analytic estimates
Number of language vars.	31	73	182	2,000	> 10,000	200	768	(various studies)
Age (r)	.56 (.55, .56)	.65 (.65, .66)	.68 (.68, .69)	.81 (.81, .81)	.83 [a]			
Gender (accuracy)	.00 (.74, .75)	.78 (.78, .79)	.82 (.81, .82)	.89 (.89, .89)	.92 [a]			
Personality								
Agreeableness (r)	.21 (.20, .22)	.26 (.25, .27)	.25 (.24, .26)	.32 (.32, .33)	.35 [b]	.33 [c]	.37 [c]	.29 (.21, .36)
Conscientiousness (r)	.26 (.26, .27)	.28 (.27, .28)	.31 (.30, .31)	.37 (.36, .37)	.37 [b]	.37 [c]	.38 [c]	.35 (.29, .42)
Extraversion (r)	.22 (.21, .23)	.30 (.29, .31)	.30 (.29, .30)	.38 (.38, .39)	.42 [b]	.37 [c]	.39 [c]	.40 (.33, .46)
Neuroticism (r)	.20 (.19, .21)	.24 (.23, .25)	.27 (.26, .27)	.34 (.33, .35)	.35 [b]	.37 [c]	.38 [c]	.33 (.27, .39)
Openness (r)	.26 (.25, .26)	.30 (.30, .31)	.33 (.32, .33)	.43 (.43, .44)	.43 [b]	.39 [c]	.44 [c]	.39 (.30, .48)
Average Personality (r)	.23	.28	.29	.37	.38	.37	.39	.35
Average Pers. Adj. R^2	.05	.08	.08	.11				

Note: For continuous outcomes, prediction performance is given by the Pearson correlation between the predicted and actual values. For gender, performance is given by classification accuracy of a penalized logistic regression model. For comparability, all language variables were extracted using DLATK (Schwartz et al., 2017). Performances for "LDA Topics, Words, Phrases" were reported in [a]Sap et al. (2014) and [b]Park et al. (2014); for vector semantic (Word2Vec) and contextual (BERT) embeddings in [c]Lynn, Balasubramanian & Schwartz (2020) disattenuated for measurement reliability (= .734). For BERT, we reported the BERT + DAN model. Meta-analytic estimates were reported in Azucar et al. (2018). Parentheses indicate 95% confidence intervals.



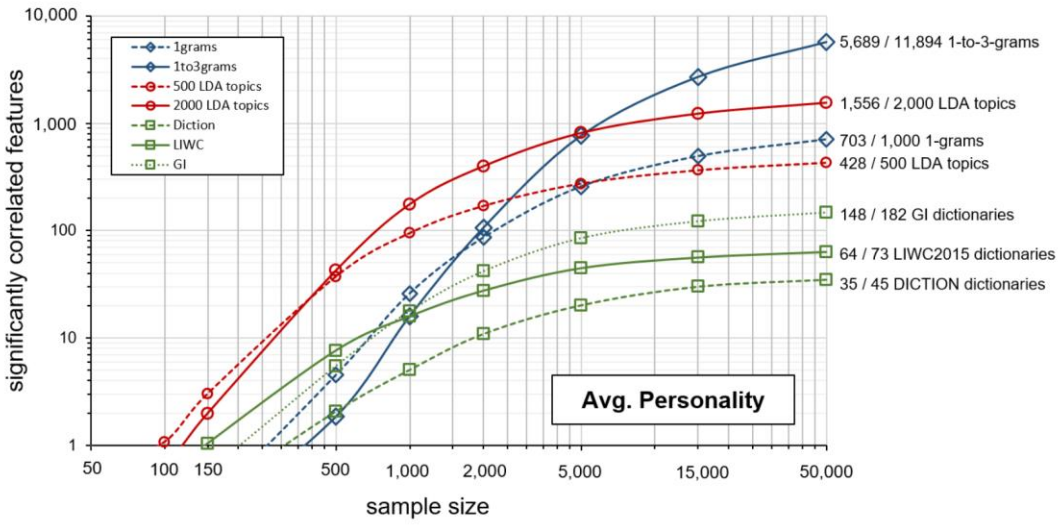


Figure 5. Average number of language features that were significantly associated with age and gender (top) and personality (bottom) as a function of sample size (log-transformed) for different feature sets. For sample sizes of 50 to 150, the significantly associated features shown are the average of 100 random draws from the overall sample (*N*=65,986); sample sizes of 500, 1,000, 5,000, 15,000 are based on 50, 20, five and three random draws, respectively. All the language of a given user was included (an average of 4,104 words). Age was controlled for gender, gender for age, and personality traits for both. Numbers of features shown are non-normalized raw counts, therefore LDA topics and the 1-to-3 grams will necessarily show higher values on the vertical axis due to having more available features.

Table 3

Sample sizes needed to observe 10 significantly associated LIWC dictionaries, 100 LDA topics or 200 1-to-3 grams for gender, age, and personality.

Thresholds of significant correlates:	Demographics		Big Five Personality					
	Gender	Age	Agr.	Con	Ext.	Neur.	Ope.	**(avg.)**
10 (out of 73) LIWC dictionaries	200	150	800	400	800	1,100	550	**750**
100 (out of 2,000) LDA topics	250	150	1,100	550	800	1,800	550	**1,000**
200 (out of 11,894) 1-to-3 grams	650	200	3,650	1,850	2,600	4,750	2,100	**3,000**

Note: All available language from users was included (an average of 4,104 words per user).

Impact of words per person

Figure 6 shows the number of significantly associated language features (after BH correction) with personality (averaged across the five traits, controlled for age and gender) as a function of different numbers of words per user for three sample sizes (N = 150, 1,000 and 5,000 users) (see Supplementary Material for each personality dimension, and for age and gender). Generally, sample size and number of words per user trade off, such that the larger the sample size, the fewer words were needed per user to reach a meaningful number of significant associations (see Table 4). Similar to the findings in the previous section, age and gender showed stronger language signal and thus fewer words per user were needed than for personality.

Specifically, for age and gender, from a sample size of N = 1,000 users a few hundred words were needed per user, depending on the choice of language variable. For LIWC and LDA topics for personality, an order of magnitude more words per user were needed – thousands of words from a sample of N = 1,000 users, or hundreds of words from a sample of N = 5,000. Finally, to reach a meaningful number (such as ~200 significant features) of 1-to-3-gram associations, thousands of words are needed from thousands of users, such as ~4,000 words from 3,000 users, as reported in Table 3.

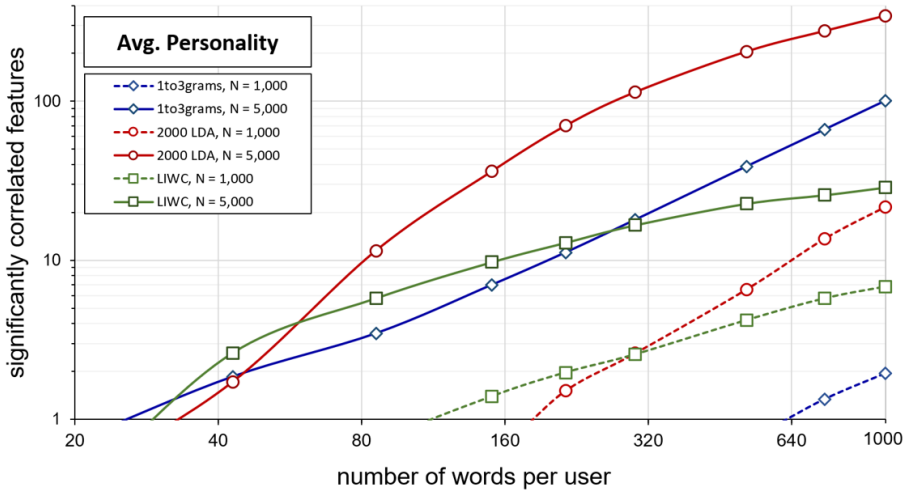


Figure 6. Average number of language features significantly associated across personality dimensions as a function of words per user (log-transformed). Associations are controlled for age and gender and given for sample sizes of N = 1,000 and 5,000, averaged across 50 and 10 random draws of users from the overall sample (*N*=65,986), respectively. Words were included from the most recent Facebook posts for a given user, in increments of whole posts (21.45 tokens per post, on average). Numbers of features shown are non-normalized raw counts, therefore LDA topics and the 1-to-3 grams will necessarily show higher values on the vertical axis due to having more features. Across all language features, no significant personality language associations were observed for a sample of N = 150. See Supplementary Materials for additional figures

Table 4

Number of words needed per user *to observe 10 significantly associated LIWC dictionaries, 100 LDA topics or 200 1-to-3 grams for demographics and personality, for sample sizes of 150, 1,000 and 5,000 users.*

	Age & Gender (avg.)			Personality (avg.)		
Sample Sizes:	N=150	N=1,000	N=5,000	N=150	N=1,000	N=5,000
10 (out of 73) LIWC dictionaries	~4,000+	90+	20+	-	~1,000 to 4,000+	170+
100 (out of 2,000) LDA topics	~4,000+	150+	40+	-	~4,000+	300+
200 (out of 11,894) 1-to-3 grams	-	750+	240+	-	-	1,000 to 4,000+

Note: For missing values, the threshold number of meaningful associations was not be reached even when including all of the users' language (an average of 4,104 words).

Closed-Vocabulary Approaches: Drivers of Prediction Errors

Closed-vocabulary programs have provided numerous insights for psychology but are also susceptible to errors. The methods compared here use a 'bag-of-words' approach, in which words are counted regardless of their context, including negation or irony. In previous work (Schwartz et al., 2013c), raters examined 100 Facebook statuses that contained words from the LIWC2007 ***positive*** and ***negative emotion*** dictionaries and rated occurrences of false positive errors. Most errors were due to lexical ambiguities (word sense and part of speech), with only 21% due to negation and 30% due to other sources. To estimate the false positive error rate of dictionaries as a measure of their specificity, human raters should rate a subset of text as to whether the occurrence of dictionary words correctly reflect the dictionary concept intended, especially if the dictionary findings are critical to the argument being made.

When using dictionaries, we have found that it is prudent to identify which words may be driving the results and consider whether the category label appropriately captures those words. To make the content of the dictionaries transparent and aid in validation, we determined the most frequent words in every dictionary used in this comparison (see Supplementary Materials). In addition, for DICTION and LIWC2007 and LIWC2015, we determined the most frequent word in the dictionary, using WordNet (Princeton University, 2010) to determine the most frequent sense of the word, and compared this word sense against the intended dictionary concept (see Supplementary Materials).

For DICTION, we found that in six dictionaries (***aggression***, ***centrality***, ***rapport***, ***exclusion***, ***liberation***, ***praise***), the most frequent word sense of the most frequent word did not match the intended dictionary concept. For example, ***liberation*** is intended to capture the maximization of individual choice and the rejection of social conventions (Hart, 2000). According to common word usage, the most frequent word "left" has the most frequent sense of "going away from a place" (Princeton University, 2010) rather than "political left", as intended by the dictionary.

For LIWC2007, we observed seven such cases (***money***, ***sadness***, ***biological processes***, ***sexual***, ***health***, ***friends***, ***time)*** (see Figure 7 for examples). For example, one of the most frequent word in the ***friends*** dictionary was *honey*, which has the most frequent sense of "a sweet yellow liquid produced by bees" (Princeton University, 2010). Of note, we found no such shortcomings in LIWC2015.

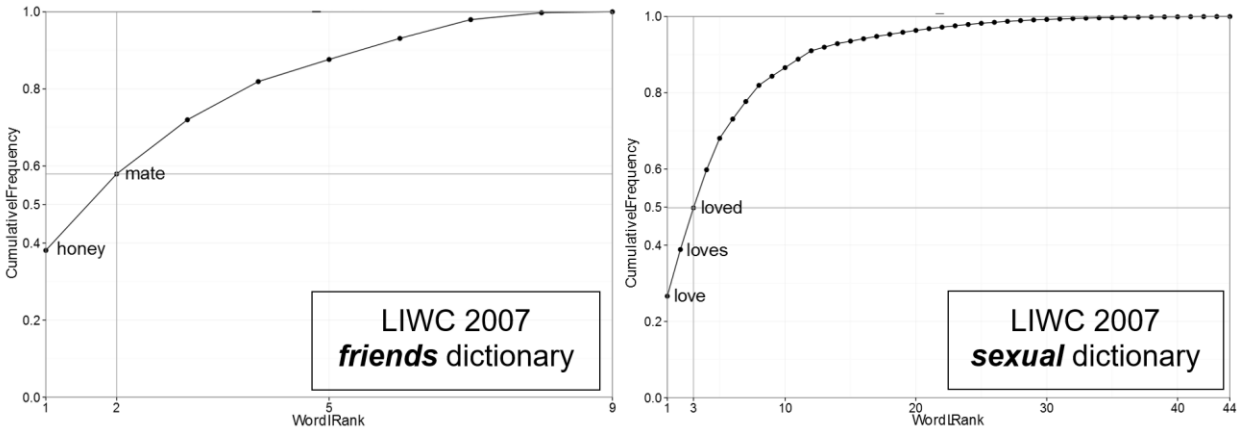


Figure 7. Cumulative frequency distributions of the LIWC 2007 *friends* (left) and *sexual (right)* dictionaries. 50% of the dictionary counts are due to two-three words, and the leading words in the dictionaries are ambiguous in word sense.

We recommend that users also manually check the most frequent words within the dictionaries being used (see Supplementary Materials on OSF)[11] as illustrated by our example. Notably, programs are increasingly adding helpful tools to help guide interpretations. For instance, LIWC2015 provides a highlighting tool ("color-code text"). For significantly correlated categories, users can use the highlights to visually identify the words that are driving the correlation. Users may thus determine if there is a mismatch in word sense or context between the dictionary and the context in which the dictionary is being applied, which may reflect specific characteristics of the population or language sample under study.

Open-Vocabulary Approaches: Choosing the Number of Topics to Extract

Table 5 shows the topics that have the word *play* among their top 10 words, across topic sets of 50, 500, and 2,000, modeled over the same five million statuses. While 50 topics failed to distinguish *ball play*, *musical play*, and *videogame play*, 500 topics successfully distinguished these contexts. The 2,000 topics distinguished different kinds of video games (i.e., military first-person shooters, real-time strategy, and action-adventure games). Finally, Figure 8 illustrates prediction accuracies using 50, 500, and 2,000 topics, modeled across varying numbers of Facebook statuses. The prediction models based on 500 or 2,000 topics were comparable and outperformed those built over 50 topics.

Table 5

Top ten words for topics that included "play" among their top 10 words for sets of 50, 500, and 2,000 topics modeled over the same 5 million Facebook statuses.

Top. Set	Occ.	Top 10 words comprising each topic
50	1	game, play, win, playing, football, team, won, games, beat, lets
500	5	guitar, play, playing, music, piano, band, bass, hero, practice, played
		game, football, play, soccer, basketball, playing, games, team, practice, baseball
		play, playing, game, ball, games, played, golf, tennis, poker, cards
		play, playing, game, games, xbox, halo, wii, video, mario, 360
		place, chuck, find, meet, play, birth, norris, interesting, babies, profile
2000	9	play, guitar, learn, piano, learning, playing, learned, lessons, songs, rules
		play, game, let's, role, sims, rules, chess, basketball, plays, poker
		play, playing, tennis, cards, wii, played, poker, ball, basketball, pool
		soccer, football, game, play, team, basketball, playing, ball, practice, field
		black, cod, ops, playing, play, mw2, modern, warfare, ps3, online
		play, playing, starcraft, warcraft, sims, ii, beta, online, nerds, nerd
		xbox, 360, play, ps3, playing, games, creed, assassin's, playstation, assassins
		words, comment, note, play, wake, jail, copy, paste, sport, fair
		games, play, playing, game, video, played, card, board, begin, playin

Note. Words suggesting playing music are highlighted in green, ball sports in blue, and videogames in yellow.

[11] See the spreadsheets for all dictionaries at https://osf.io/qtajf/.

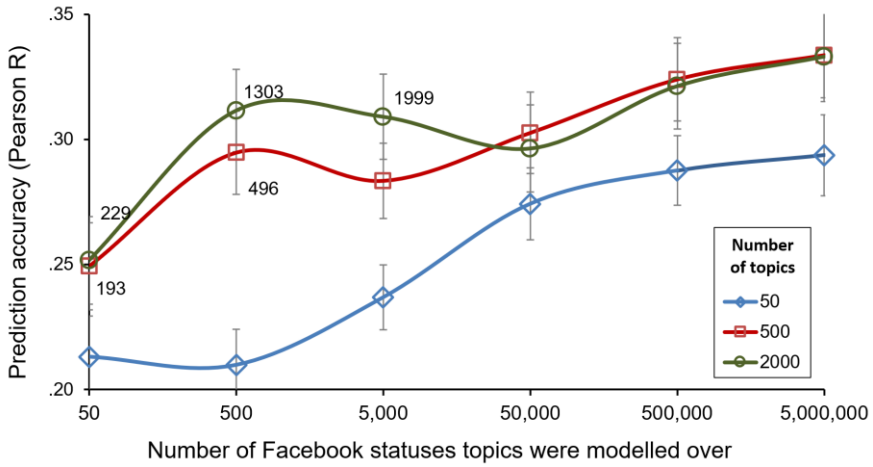


Figure 8. Prediction accuracies across 65,896 users and 12.7 million Facebook statuses obtained using 50, 500, and 2,000 topics, modeled across 50 to five million Facebook statuses. Cross-validated ridge-regression prediction accuracies were averaged across the five traits; error bars give the standard error of the mean. When the number of topics to be modeled was close to or exceeded the number of statuses to be modeled over, the modeling algorithm created fewer topics; in those case the actual number of topics modeled is noted.

Discussion

There is a raft of remarkable work in language analysis in fields related to psychology, including style matching (Gonzales, Hancock, & Pennebaker, 2010; Ireland et al., 2011; Taylor & Thomas, 2008); how power differentials amongst participants and engagement in the community affect language (Danescu-Niculescu-Mizil et al., 2012; Danescu-Niculescu-Mizil et al., 2013), understanding personal values (Boyd et al., 2015) and what makes content go viral (Berger & Milkman, 2012); and identifying emotions (Bollen, Mao, & Pepe, 2011; Strapparava & Mihalcea, 2008) and psychological traits (Guntuku et al., 2017; Mitchell, Hollingshead, & Coppersmith, 2015; Park et al., 2015; Sagi & Dehghani, 2014) in textual data. These studies (among many others) point to the potential of what is possible by incorporating language into psychological research. The psychological literature on text-based data thus far has relied almost entirely on closed-vocabulary programs, which were carefully developed for specific purposes. Open-vocabulary approaches extend these traditional programs, providing data-driven approaches for making predictions and gaining insights. As Pennebaker and colleagues foresaw in 2003, "for researchers interested in learning what people say--as opposed to how they say it--we recommend this new analytic approach" (p. 571).

Psychological research has evolved considerably over the past decade, expanding the questions that can be asked, the phenomena that can be studied, and the methods that can be used. Experimental studies, and recommendations around sample size and significance developed in a period where access was limited to local environments and calculations occurred by hand. Similarly, closed-vocabulary approaches originated at a time when very few large-scale correlational studies existed, and limited amounts of text were recorded in experimental contexts in which qualitative information was hard to capture. Social media and other online sources now make large amounts of textual data readily accessible, and automatic approaches allow for the efficient processing of large-scale analyses. Our review suggests that there is benefit in carefully using closed-vocabulary approaches for some questions, such as identifying how people think, or

testing specific hypotheses, but points to the benefit of increasingly incorporating open-vocabulary approaches to understand what people specifically think about, and how that drives subsequent thoughts, emotions, and behaviors in everyday life.

To provide guidance to the effective application of possible approaches to text analysis, this synthesis quantitatively compared five closed- and open-vocabulary methods across 13 million Facebook status updates from over 65,000 users. Open-vocabulary results were congruent with, but conceptually more specific, than closed-vocabulary results, pointing to specific behaviors and emotions not captured by the dictionaries. For example, while male language was associated with ***hostility*** and ***aggression*** dictionaries, LDA topics revealed these associations to be due to references to competition, political debate, and sports.

Cross-validated machine learning prediction models indicated that the 2,000 LDA topics captured the most demographic- and personality-related variance in language, followed by LIWC2015 and GI, which captured roughly equal amounts of variance. The language results expand and update previous studies on the association of language with age (e.g., Kern et al., 2014b; Pennebaker & Stone, 2003; Schwartz et al., 2013b), gender (e.g., Newman et al., 2008; Schwartz et al., 2013b), and personality (Kern et al., 2014a; Schwartz et al., 2013b; Yarkoni, 2010). GI, DICTION, and LIWC2015 overlap in their coverage of pronouns and concepts, including positive and negative emotion, complex language suggestive of higher cognition, economic and fiscal concerns, and social and family relationships. The dictionaries that distinguished positive and negative emotions were among those most associated with female gender, older age, higher levels of agreeableness, conscientiousness, and extraversion, and lower levels of neuroticism.

While effect sizes varied by approach, our results illustrate that the content of what people write about in everyday life is indeed related to who they are as a person, including their age, gender, and personality. Various studies have attempted to show this, using closed-vocabulary approaches (e.g., Gill, Nowson, & Oberlander, 2009; Golbeck, Robles, & Turner, 2011; Sumner, Byers, & Shearing, 2011). Similar to previous work (Iacobelli, Gill, Nowson, & Oberlander, 2011; Schwartz et al., 2013b), the open-vocabulary prediction models outperformed dictionary-based prediction models, suggesting that the larger number of open-vocabulary features capture more of the personality-related variance in the language data. This suggests that open-vocabulary methods are particularly suited for capturing the nuances of everyday psychological processes. This is fundamentally different from what the closed-vocabulary approaches were initially intended for, such as coding reflective essays (which LIWC is well suited for) or analyzing presidential speeches (the purpose for which DICTION was created).

Recommendations for Researchers

Based on our review, we provide recommendations for research in this area, including consideration of the approach, using closed- and open-vocabulary approaches, and sample size.

Choosing an approach. Closed-vocabulary programs have been instrumental in providing tools for quantifying text-based information. They have several properties that make them desirable: a contained set of dictionaries yields a relatively parsimonious quantitative representation of language content; as the dictionaries are the same across studies, the results are comparable; and they are well-suited to reliably capture patterns among function words that do not suffer from word sense ambiguities. Validated dictionaries can be suitable for testing specific hypotheses. But dictionary-based approaches also have sources of potential errors, so care should be taken when relying on single dictionary associations.

Open-vocabulary approaches yield more specific language insights into why associations

may occur, which are useful for generating new hypotheses and understanding underlying processes. They can unpack the closed-vocabulary results. They also capture more construct-related variance in the language (i.e., have higher predictive power). Open-vocabulary approaches create transparent units of language, and results can be shortlisted, filtered for uninformative duplicates, and visualized for inspection as a list or word cloud, yielding intuitive summaries of what language most distinguishes a characteristic. However, word, phrase, topic or embedding extraction an be harder to implement and require more expertise. Sample size and number of words per user also needs to be appropriate, and the number of topics to be extracted needs to be considered.

Ideally, closed- and open-vocabulary approaches should be combined. Even when conducting open-vocabulary analyses, a set of dictionaries allows the researcher to quickly get a sense of the language correlates of a given trait before examining a potentially large number of topic correlations in more detail. In this way, closed-vocabulary correlations can help the researcher see the broad patterns, which the fine-grained open-vocabulary approaches can then unpack. Over 15 years ago, Pennebaker, Mehl, and Niederhoffer (2003) foresaw that word count approaches based on dictionaries defined by the researcher would eventually be complemented by methods from artificial intelligence. This has now become a reality, with considerable benefit in considering how the two can be used together to provide the greatest insights into psychological processes.

Sample size and words per user considerations. One advantage of dictionary-based methods is their relatively smaller number of language features (i.e., dictionaries), compared to the very large number of words, phrases, and topics used in DLA. This points to the different discipline intentions for which textual analyses typically are performed. In computer science, the goal often is accurate prediction and theory-free exploration, such that a large number of features is preferable. In psychology, the goal often is understanding mechanisms and testing theory, such that a small number of theoretically relevant variables is preferable. Depending on the purpose, sample size, and textual data size, LIWC or LDA topics may provide greater insights or be more useful in the scientific process.

In terms of sample sizes, if thousands of words are available from a given user, as is the case with histories of Facebook statuses, we found that for both demographics and personality, language profiles with sufficient nuance were observed with similar sample sizes for the LDA topics and the LIWC2015 dictionaries (N ~ 250 vs. 200 for demographics, N ~ 1,000 vs. 750 for personality; see Table 3). This may seem surprising, given that 2,000 LDA topics are more numerous than 73 LIWC dictionaries. Substantially more participants are needed for word and phrase correlations (N ~ 650 for demographics and N ~ 3,000 for personality). Regardless of the purpose, to avoid the risk of spurious findings, the customary significance thresholds should be corrected for the number of language features being tested, and indications of significance should be used only as a heuristic for potentially meaningful results.

In terms of textual size, the sample size and the number of words per user trade off against one another in terms of statistical power, such that for larger sample sizes, fewer words per users are needed, and, reversely, if more words per user are available, smaller sample sizes may be adequate. For example, nuanced language profiles for age and gender for LIWC and LDA topics could be observed with as little of 20-40 words per user for a sample of N = 5,000 users, while for a sample of N = 150, thousands of words per user were required (see Table 4 for details). Generally, more textual data is required to explore the language of personality than for demographics. As a rule of thumb, for personality, for both LIWC and topics, for a sample of

order N = 1,000, thousands of words are needed from a user. For a sample of N ~ 5,000, hundreds of words may suffice. As reported above, for nuanced words and phrase correlations, substantially more textual data is required -- thousands of users have to provide thousands of words. (For comparison, an average Facebook post in the study data set is about 21 words long, and an average Tweet is around 15 words.) Of note, these considerations cover exploratory language analyses – in experimental research, specific language variables may be hypothesized to change as a result of experimental condition, and accordingly, the thresholds given here may overestimate the amount of textual data that is required (see Supplementary Materials (https://osf.io/h4y56) for more detailed figures about when first significant language correlations emerge).

Dictionary considerations. Among the closed-vocabulary approaches, LIWC has been used most frequently for psychological text analysis. The 2015 version clearly improves upon the 2007 version, and its 73 dictionaries appear to be a strong contender in terms of effectively balancing exhaustiveness and parsimony. GI was ahead of its time and provides dictionaries on par in coverage (but not parsimony) with LIWC2015, and its dictionaries are free for non-commercial use. DICTION covers fewer language concepts, and its method of combining multiple dictionaries into master variables is not recommended, as the results can be hard to interpret, especially if any of the underlying dictionary associations are misleading. Most (but not all) of the dictionaries provide acceptable measures of their intended constructs. Whereas GI and LIWC were developed more broadly to capture psychological and sociological phenomenon, DICTION was developed specifically for use with political communication. The particulars of the research domain, theories, assumptions, and design of both the dictionaries and the context in which the dictionaries will be applied should be considered, and, if in doubt, validated.

Because of the Zipfian distribution of language, the overall frequencies of dictionaries are often determined by a few highly frequent words. Therefore, it is useful to first consider whether the most frequent word sense for a given dictionary's most frequent words correctly captures the dictionary concept. Better yet, dictionaries should be validated for a given language sample, particularly when the validity of a given dictionary is the basis for theoretical inference, or when a dictionary is applied to language contexts different from those for which it was designed (see Grimmer & Stewart, 2013 for the validation process, and Eichstaedt et al., 2015, Schwartz et al., 2013b, and Sun et al., 2019 for examples).

Topic model considerations. Topics that arise through LDA have the advantage of keeping individual words within their context. A cluster of words in a topic can be a more dependable unit of analysis than single word associations, or dictionaries that are dominated by ambiguous, highly frequent words. Creating topics based on a given language corpus is also an efficient way of summarizing the themes mentioned in the corpus.

Generally, the larger the corpus, the more coherent and fine-grained topic models can be constructed. As a lower limit, a customary rule of thumb suggests that one should have at least 50 documents for every LDA topic being modelled, in the same way that a sufficient sample size is needed to factor analyze a set of items (see Kern et al., 2016 for considerations regarding the amount of linguistic and outcome data needed to generate meaningful results). Notably, it is not necessary to develop the topics on the same language dataset to which they are applied. This creates the possibility of creating topic models on a larger language sample which contains more semantic information to inform the modeling process, and then applying the topics to a smaller study sample. This mirrors the “off-the-shelf” use of dictionaries, but topics are driven by the data rather than by theory. Using the same set of topics across multiple studies and datasets can

also allow researchers to compare topic results across datasets. Future work might establish a consistent set of data-driven topics that can be used across studies within a particular domain, similar to how the theoretically-derived dictionaries have been used to date.

If one has sufficient data, our analysis suggests that the number of topics needed depends on the goal of the study. If the goal is accurate predictions, one ought to err on the side of modeling more rather than fewer topics. Overall, in large social media data sets with millions of documents, we have found 500-2,000 topics to provide the right level of nuance, and visualizing the most correlated topics to yield a general view of what users are writing about. Larger numbers of topics (in the thousands) will contain many near duplicates and may lower the ability to establish exploratory language profiles when correcting for multiple comparisons. If the language domain, the study context or the sample size is narrower, modeling a smaller number of topics maybe appropriate. For example, we found 200 topics to provide the right level of nuance across a sample of about 1,000 Facebook users with 1 million Facebook statuses recruited in a medical context to study depression (Eichstaedt et al., 2018).

A large literature discusses methods to automatically determine the optimal number of topics to extract across different kinds of language data, including methods that consider statistical perplexity or rates of perplexity change to determine the optimal number of topics (e.g., Zhao et al., 2015). However, other studies have shown these statistical measures (and other measures such as prediction performance) to be poor predictors of human judgments of topic quality and semantic coherence (e.g., Chang et al., 2009). Thus, at this time, we recommend avoiding fully automated models, and manually inspecting topic quality.

Of note, many function words are not suitably captured in the topic modeling process. Due to their syntactic omnipresence in the language across different contexts, they would appear in most topics, such that they are routinely excluded when topics are modeled. We therefore recommend adding the 200 most frequent words (or ***function word*** dictionaries) as additional language variables to analyses that would otherwise be limited only to LDA topics.

Resources and Tools

Part of LIWC's success has been the ease of use of the program. While many packages exist to perform topic modeling (such as Mallet; McCallum, 2002), none of them currently is as easy to use as LIWC. However, other methods are also becoming easier to use. All of the analyses in this comparison can be carried out using the open-source DLATK Python code base (Schwartz et al., 2017; see dlatk.wwbp.org for a number of tutorials). DLA can also be carried out online (http://lexhub.org). In addition, in the Supplementary Materials, we share the 500 and 2,000 topics in the form of "weighted dictionaries" that can be used by other text analysis programs,[12] as well as the GI dictionaries that capture as much trait-related variance as LIWC, but are free for non-commercial use (see https://osf.io/h4y56).

Limitations

While this review compares three closed-vocabulary and two open-vocabulary approaches, it does not address the ways in which supervised machine learning methods might augment or even replace annotation by humans (for a review, see Grimmer & Stewart, 2013), or how dictionaries can be improved using data-driven approaches (e.g., Sap et al., 2014, Schwartz et al. 2013c). We did not discuss the many emerging algorithms to create topic models that take user attributes into account. We also omitted a discussion of how dimensionality reduction techniques can be combined to create more parsimonious representations of the language space (e.g., multi-level LDA, or a combination of LDA topic modeling with matrix factorization

[12] Unfortunately, LIWC2015 does not support weighted dictionaries.

techniques). These methods are yet to be introduced to psychological research and are areas that should be explored in the future, especially in terms of their suitability and applicability.

Opportunities on the Horizon

We have reviewed several existing closed- and open-vocabulary approaches for automated text analysis. As approaches from computational linguistics in psychology are fairly new, these approaches are simply the beginning of what may be possible. We end with consideration of what may be on the horizon.

Word and contextual embedding models are just beginning to be used for psychological insight. In this review, we have discussed Differential Language Analysis which uses purely lexical features with no regard for context. In principle, the deep contextual knowledge that is encoded in contextual approaches (such as BERT) is ripe for extraction to study differences between people and cultures. Future work may address how this knowledge can be meaningfully extracted and distilled in a way that informs psychological theory.

So far, semantic distances between concepts in embedding spaces have been used to measure the associations or similarity that these concepts hold globally in human minds (e.g., Bhatia, 2017) – but these methods have not yet been used to study the differences between human minds. It is conceivable that training different semantic representations for different personality profiles may give us a glimpse into individual differences in knowledge and concept representations.[13] Further, in experimental or intervention research, training different embedding spaces across the writings of different treatment conditions may make the cognitive impact of psychological interventions measurable as relative differences or changes in semantic distances.

More generally, in regard to experimental research, throughout this manuscript we have observed that off-the-shelf dictionaries may often be suitable to test specific hypotheses. However, in situations where such training data is available, supervised open-vocabulary prediction models can be trained to measure psychological states and traits from text, in the same way that personality was predicted in this review. Language-based prediction models use the entirety of the vocabulary, and can provide assessment of variables of theoretical interest with more sensitivity than through closed-vocabulary approaches. An increasing number of such language-based assessment models are "on the shelf" (e.g., temporal orientation: Park et al., 2015b, valence/arousal: Eichstaedt & Weidman, 2020, or empathy: Abdul-Mageed et al., 2017). The "revolution" of contextual embedding methods in NLP will lead to increasingly accurate text-based measurement models in psychology that are ripe for use in large scale experimental contexts, were scalable psychological measurement of populations may be desired.

Conclusion

Written language, whether hand-written or typed on a computer or smart device, is a core way that humans communicate, conveying thoughts, emotions, and traces of themselves to others. The rapid growth and availability of large amounts of digitized textual data, combined with programs developed within the social and computer sciences, have created the opportunity to study psychological processes as they happen in everyday life, at a scale never before possible.

This potential must be matched with careful consideration of the purpose of the study, the data available, and the analytic approaches used. Just as other areas of psychology have found that constructs of interest are best measured through a combination of approaches, our analysis suggests that the methods compared here provide complementary lenses. The closed- and open-vocabulary findings are surprisingly consistent. Each one has strengths and weaknesses, but the

[13] Bhatia (2017) also remarked on this promising direction.

combination provides the clearest view of language correlates of psychological constructs. Dictionaries of function words are powerful markers of underlying cognitive and attentional psychological processes, and together with positive and negative emotion dictionaries are often among the most distinguishing markers for personality and demographic traits. Topic models--either modeled on the same corpus or imported from a larger one—produce more fine-grained, contextually-embedded, transparent units of analysis than do dictionaries, and allow for the discovery of specific emotions, thoughts, and behaviors. Closed-vocabulary approaches can be rigid, while open-vocabulary approaches can be sensitive to idiosyncrasies of the dataset and the modeler's choices about parameters. Closed approaches are more reproducible but inflexible, whereas open approaches are more flexible but can vary across datasets.

The largest datasets of our digital era are textual in nature. While computational approaches may prevail, both closed and open-vocabulary approaches are needed to allow psychologists to test hypotheses and to discover new ones. Closed-vocabulary approaches provide a powerful way to study *how* people think, while open-vocabulary approaches elucidate *what* people think about. Together, these approaches allow us to study psychological processes as they occur in everyday life in the largest longitudinal, cross-sectional, and cross-cultural study in human history.

References

Abdul-Mageed, M., Buffone, A., Peng, H., Eichstaedt, J. C., & Ungar, L. H. (2017). Recognizing Pathogenic Empathy in Social Media. *In Proceedings of the Twelfth International AAAI Conference on Web and Social Media* (ICWSM). pp. 448-451.

Alderson, J. C. (2007). Judging the frequency of English words. *Applied Linguistics, 28*, 383-409.

Almodaresi, F., Ungar, L., Kulkarni, V., Zakeri M., Giorgi, S. & Schwartz, H. A. (2017). On the distribution of lexical features in social media. *Annual Meeting of the Association for Computational Linguistics.* Vancouver, Canada.

Atkins, D. C., Rubin, T. N., Steyvers, M., Doeden, M. A., Baucom, B. R., & Christensen, A. (2012). Topic models: A novel method for modeling couple and family text data. *Journal of Family Psychology*, *26*(5), 816–827.

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences, 124*, 150-159.

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, 57*, 289-300.

Berger, J., & Milkman, K. L. (2012). What makes online content viral?. *Journal of Marketing Research*, *49*(2), 192-205.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Bollen, J., Mao, H., & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *ICWSM*, *11*, 450-453.

Boyd, R. L., & Pennebaker, J. W. (2015). A way with words: Using language for psychological science in the modern era. *Consumer Psychology in a Social media World*, 222-236.

Boyd, R. L., Wilson, S. R., Pennebaker, J. W., Kosinski, M., Stillwell, D. J., & Mihalcea, R. (2015, April). Values in words: Using language to evaluate and understand personal values. In *ICWSM* (pp. 31-40).

Campbell, R. S., & Pennebaker, J. W. (2003). The secret life of pronouns: Flexibility in writing style and physical health. *Psychological Science*, *14*, 60–65.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems* (pp. 288-296).

Chung, C., & Pennebaker, J. (2007). The psychological functions of function words. In *Social Communication* (pp. 343–359). Taylor and Francis. doi:10.4324/9780203837702

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.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012, April). Echoes of power:

Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web* (pp. 699-708). ACM.

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013, May). No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 307-318). ACM.

Deerwester, S. C., Dumais, S. T., Furnas, G. W., Harshman, R. A., Landauer, T. K., Lochbaum, K. E., and Streeter, L. (1988). Computer information retrieval using latent semantic structure: US Patent 4,839,853..

Deerwester, S. C., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Schwartz, H. A., & Gomez, F. (2008, August). Acquiring knowledge from the web to be used as selectors for noun sense disambiguation. In CoNLL 2008: *Proceedings of the Twelfth Conference on Computational Natural Language Learning* (pp. 105-112).

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., … Seligman, M. E. P. (2015). Psychological language on twitter predicts county-level heart disease mortality. *Psychological Science, 26*, 159-169.

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Daniel Preotiuc-Pietro, D., Asch, D. A., & Schwartz, H. A. (2018) Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences., 115 (44), 11203-11208

Eichstaedt, J. C.& Weidman, A. (2020, in press) Tracking Fluctuations in Psychological States: A Case Study of Weekly Emotion Using Social Media Language. *European Journal of Personality*

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008) Liblinear: A library for large linear classification. *Journal of Machine Learning Research, 9*, 1871-1874.

Francis, M. E., & Pennebaker, J. W. (1992). Putting stress into words: The impact of writing on physiological, absentee, and self-reported emotional well-being measures. *American Journal of Health Promotion*, *6*, 280-287.

Francis, M. E., & Pennebaker, J. W. (1993). *LIWC: Linguistic inquiry and word count*. Dallas, TX: Southern Methodist University.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*, 137-144.

Gilbert, E. (2012, February). Phrases that signal workplace hierarchy. *In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 1037-1046). ACM.

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.

Glaser, B. G., & Strauss, A. L. (1967) *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine.

Gleser, G. C., Gottschalk, L. A., & Springer, K. J. (1961). An Anxiety Scale Applicable to Verbal Samples. *Arch Gen Psychiatry*, *5*(6), 593–605.

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., 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.

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint* arXiv:1402.3722.

Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2010). Language style matching as a predictor of social dynamics in small groups. *Communication Research, 37*(1), 3-19.

Gottschalk, L. A., & Gleser, G. C. (1969). *The measurement of psychological states through the content analysis of verbal behavior*. Univ of California Press.

Gottschalk, L. a., & Bechtel, R. (1995). Computerized measurement of the content analysis of natural language for use in biomedical and neuropsychiatric research. *Computer Methods and Programs in Biomedicine, 47*(2), 123–130.

Gottschalk, L. A., & Bechtel, R. J. (2000). PCAD 2000. *Psychiatric Content Analysis and Diagnosis: GB software.* Corona del Mar, CA, 4607.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*, 267–297.

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences, 18*, 43-49.

Hart, R. P. (1984). *Verbal style and the presidency: A computer-based analysis*. Academic Pr.

Hart, R. (2001). Redeveloping Diction: Theoretical considerations. In *Theory, Method, and Practice in Computer Content Analysis* (pp. 43-60). Westport, CT: Ablex Publishing.

Hart, R. P. (2000). *Diction 5.0 User's Manual*. Austin, TX: Digitext, Inc.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*, 55-67.

Holsti, O. R., Brody, R. A., & North, R. C. (1964). Measuring affect and action in international reaction models: Empircal materials from the 1962 Cuban crisis. *Journal of Peace Research, 1*, 170-189.

Iacobelli, F., Gill, A. J., Nowson, S., & Oberlander, J. (2011). Large scale personality classification of bloggers. In *Affective Computing and Intelligent Interaction* (pp. 568-577). Springer Berlin Heidelberg.

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*, *22*(1), 39-44.

Inquirer Home Page. (2002, September 12). Retrieved from http://www.wjh.harvard.edu/~inquirer/

Iliev, R., Dehghani, M., & Sagi, E. (2014). Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition*, 1–26.

Jaidka, J., Giorgi, S. Schwartz, H. A., Kern, M. L., Ungar, L. H., Eichstaedt, J. C. (2020). Estimating geographic subjective well-being from Twitter: a comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*.

Jurafsky, D., Martin, J. H. (2019) Speech and Language Processing.

rights reserved. Draft of October 2, 2019.

Kelly, E. F., & Stone, P. J. (1975). *Computer recognition of English word senses* (Vol. 13). North-Holland.

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Dziurzynski, L., Ungar, L. H., Stillwell, D. J., Kosinski, M., Ramones, S. M., & Seligman, M. E. (2014a). The online social self: An open vocabulary approach to personality. *Assessment, 21*, 158-169.

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Park, G., Ungar, L. H., Stillwell, D. J., … & Seligman, M. E. P. (2014b). From "sooo excited!!!" to "so proud": Using language to study development. *Developmental Psychology, 50,* 178-188.

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological Methods*, 21, 507-525.

Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In Proceedings of NAACL-HLT* (pp. 4171-4186).

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist, 70*, 543-556.

Kosinski, M., & Stillwell, D. (2012). MyPersonality project. Available from http://www.mypersonality.org/wiki/.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America, 110*(15), 5802-5805.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240.

Lasswell, H. D., & Kaplan, A. (1950). *Power and society: A framework for political inquiry*. Transaction Publishers.

Lasswell, H. D., & Namenwirth, J. Z. (1969). *The Lasswell value dictionary*. New Haven.

Leacock, C., Towell, G., & Voorhees, E. M. (1993). Towards building contextual representations of word senses using statistical models. In *Acquisition of Lexical Knowledge from Text*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Loughran, T., & Mcdonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance, 66,* 35-65.

Lynn, V., Balasubramanian, N., & Schwartz, H.A. (2020). Message-Level Attention for User

Personality Modeling. In ACL-2020: *Proceedings of the Association for Computational Linguistics*.

Martindale, C. (1973). An experimental simulation of literary change. *Journal of Personality and Social Psychology*, *25*(3), 319–326.

McCallum, A. K. (2002). *Mallet: A machine learning for language toolkit* [computer software]. Retrieved from http://mallet.cs.umass.edu

McClelland, D. C. (1961). *Achieving society*. Simon and Schuster.

Mehl, M. R. (2006). Quantitative text analysis. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 141-156). American Psychological Association.

Mergenthaler, E., & Bucci, W. (1999). Linking verbal and non-verbal representations: Computer analysis of referential activity. *The British Journal of Medical Psychology*, *72*, 339–354.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. A. (2015). U.S. Patent No. 9,037,464. Washington, DC: U.S. Patent and Trademark Office.

Mitchell, M., Hollingshead, K., & Coppersmith, G. (2015). Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 11-20).

Morgan, C. D., & Murray, H. A. (1935). A method for investigating fantasies: The thematic apperception test. *Archives of Neurology and Psychiatry*, *34*, 289–306.

Murray, H. A. (1938). *Explorations in personality.* New York: Oxford University Press.

Murray, H. A. (1943). *Thematic Apperception Test manual*. Cambridge, MA: Harvard University Press.

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

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage Publications.

Osgood, S., Suci, G. J., & Tannenbaum, P. H. (1957). The measurement of meaning. *Urbana: University of Illinois Press*.

Osgood, C. E. (1963). On understanding and creating sentences. *American Psychologist, 18*, 735-751.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*, 934-952.

Park, G., Schwartz, H. A., Sap, M., Kern, M. L., Weingarten, E., Eichstaedt, J. C., Berger, J., Stillwell, D. J., Kosinski, M., Ungar, L. H. & Seligman, M. E. (2015b). Living in the Past, Present, and Future: Measuring Temporal Orientation with Language. Journal of Personality.

Park, G., Yaden, D. B., Schwartz, H. A., Kern, M. L., Eichstaedt, J. C., Kosinski, M., …, &

Seligman, M. E. P. (2016). Women are warmer but no less assertive than men: Gender and language on Facebook. *PLoS ONE*, *11*(5), e0155885.

Parrigon, S., Woo, S. E., Tay, L., & Wang, T. (2017). CAPTION-ing the situation: A lexically-derived taxonomy of psychological situation characteristics. *Journal of personality and social psychology*, *112*(4), 642.

Pennebaker, J. W. (2011). The secret life of pronouns. *New Scientist*, *211*(2828), 42-45.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic inquiry and word count: LIWC* [Computer software]. Austin, TX: University of Texas at Austin.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count (LIWC): A computerized text analysis program*. Mahwah, NJ.

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., & Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, *85*, 291-301.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing* (EMNLP) (pp. 1532-1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In Proceedings of NAACL-HLT (pp. 2227-2237).

Peters, M., Neumann, M., Zettlemoyer, L., & Yih, W. T. (2018). Dissecting Contextual Word Embeddings: Architecture and Representation. In Proceedings of the 2018 *Conference on Empirical Methods in Natural Language Processing* (pp. 1499-1509).

Pierce, J. (1980). *An introduction to information theory: Symbols, signals & noise* (2nd, rev. ed.). New York: Dover Publications.

Pietraszkiewicz, A., Formanowicz, M., Sendén, M. G., Boyd, R. L., Sikström, S., & Sczesny, S. (2019). The big two dictionaries: Capturing agency and communion in natural language. *European Journal of Social Psychology, 49*, 871-887.

Potts, C. (2011). *Happyfuntokenizer* (Version 10). [Computer software]. Retrieved from http://sentiment.christopherpotts.net/code-data/happyfuntokenizing.py

Princeton University (2010). *About WordNet*. Retrieved from: https://wordnet.princeton.edu

Rorschach, H. (1942). *Psychodiagnostics* (6th ed.). New York: Grune and Stratton.

Sagi, E., & Dehghani, M. (2014). Measuring moral rhetoric in text. *Social Science Computer Review*, *32*(2), 132-144.

Sap, M., Park, G., Eichstaedt, J. C., Kern, M. L., Stillwell, D. J., Kosinki, M., Ungar, L. H., & Schwartz, H. A. (2014). Developing age and gender predictive lexica over social media. *Conference on Empirical Methods in Natural Language Processing (EMLNP)*. Doha, Qatar.

Schwartz, H. A., & Gomez, F. (2008, August). Acquiring knowledge from the web to be used as selectors for noun sense disambiguation. In CoNLL 2008: *Proceedings of the Twelfth*

Conference on Computational Natural Language Learning (pp. 105-112).

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Agrawal, M., Park, G. J., … Lucas, R. E. (2013a). Characterizing geographic variation in well-being using tweets. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM)*. Boston, MA.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … Ungar, L. H. (2013b). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One*, *8*(9), e73791. doi:10.1371/journal.pone.0073791

Schwartz, H. A., Eichstaedt, J. C., Dziurzynski, L., Kern, M. L., Blanco, E., Ramones, S., Seligman, M. E. P., & Ungar, L. H. (2013c). Choosing the right words: Characterizing and reducing error of the word count approach. In **SEM-2013: Second Joint Conference on Lexical and Computational Semantics*.

Schwartz, H. A., Giorgi, S., Sap, M., Crutchley, P., Eichstaedt, J. C., & Ungar, L., H. (2017). DLATK: Differential language analysis ToolKit. In P*roceedings of the Conference on Empirical Methods in Natural Language* Processing: System Demonstrations (pp. 55-60).

Schwartz, H. A., & Ungar, L. H. (2015). Data-driven content analysis of social media: A systematic overview of automated methods. *The Annals of the American Academic of Political and Social Science, 659,* 78-94.

Smith, C. P. (Ed.). (1992). *Motivation and personality: Handbook of thematic content analysis*. Cambridge University Press.

Stone, P. J., Bales, R. F., Namenwirth, J. Z., & Ogilvie, D. M. (1962). The General Inquirer: A computer system for content analysis and retrieval based on the sentence as unit of information. *Computers in Behavioral Science, 7,* 484–498.

Stone, P., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1968). The General Inquirer: A computer approach to content analysis. *Journal of Regional Science,* 8, 113-116.

Strapparava, C., & Mihalcea, R. (2008, March). Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing* (pp. 1556-1560). ACM.

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.

Sun, J., Schwartz, H. A., Son, Y., Kern, M. L., Vazire, S. (in press). The language of well-being: tracking fluctuations in emotion experience through everyday speech. *Journal of Personality and Social Psychology*.

Tausczik, Y. R., & 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, P. J., & Thomas, S. (2008). Linguistic style matching and negotiation outcome. *Negotiation and Conflict Management Research, 1*, 263-281.

Weber, R. P. (1984). Computer-aided content analysis: A short primer. *Qualitative sociology*, *7*, 126-147.

Weber, R.P. (Ed.). (1990). *Basic content analysis*. Sage.

Wolfe, M. B., & Goldman, S. R. (2003). Use of latent semantic analysis for predicting psychological phenomena: Two issues and proposed solutions. *Behavior Research Methods*,

Instruments, & Computers, 35, 22-31.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems* (pp. 5754-5764).

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.

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015, December). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC bioinformatics, 16*, p. S8