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

Johannes C. Eichstaedt1,2, Margaret L. Kern3, David B. Yaden4, H. A. Schwartz5, Salvatore Giorgi6, Gregory Park6, Courtney A. Hagan6, Victoria A. Tobolsky6, Laura K. Smith6, Anneke Buffone6, Jonathan Iwry6, Martin E. P. Seligman6, and Lyle H. Ungar6

1 Department of Psychology, Stanford University
2 Institute for Human-Centered A.I., Stanford University
3 Melbourne Graduate School of Education, The University of Melbourne
4 Department of Psychiatry and Behavioral Sciences, Johns Hopkins Medicine
5 Department of Computer Science, Stony Brook University
6 Department of Psychology, University of Pennsylvania

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 LIWC2015 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.

Translational 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 quantitatively 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 thank Jordan Carpenter, Daniel Preotuci-Pietro, and Shriniidhi Kowshika Lakshmikanth for their help on the project, Michal Kosinski and David J. Stillwell for providing access to the MyPersonality dataset, and Molly Ireland, Jamie Pennebaker, and Ryan L. Boyd for their insightful comments on the manuscript. Martin E. P. Seligman and Lyle H. Ungar contributed equally. Correspondence concerning this article should be addressed to Johannes C. Eichstaedt, Department of Psychology, Stanford University, 450 Jane Stanford Way Building 420, Stanford, CA 94305, United States, or to Margaret L. Kern, Melbourne Graduate School of Education, The University of Melbourne, 100 Leicester Street, Level 2, Parkville, VIC 3010, Australia. Email: johannes.stanford@gmail.com or peggy.kern@unimelb.edu.au
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

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 that 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” (Iliev et al., 2015, p. 21). This ongoing 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 networked devices has become a part of an emerging longitudinal, cross-sectional, and cross-cultural study (Iliev et al., 2015, p. 21).

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” might be part of a positive emotions dictionary). The dictionaries have been incorporated into computer programs that allow text to be automatically scanned, count how often words from each dictionary occur, and output 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 et al., 2001). Still, the programs have been applied across a diverse range of contexts.

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 et al., 2015), and Latent Dirichlet Allocation (LDA; Blei et al., 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 et al., 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 widely used approaches to textual analysis across a number of fields, and it is only a question of time before it becomes 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 widespread 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. Whereas 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 an 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 timely orientation to computational methods of linguistic analysis, based on an “apples to apples” comparison for the prominent methods since their introduction in the 1960s, using a widely used dataset. 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 and instructed human raters on 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 1970s with the growth of qualitative methodologies such as grounded theory (Glaser & Strauss, 1967). Additional qualitative coding systems have been developed over subsequent 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 (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 within the English language. The dictionary relative frequencies can be compared across texts and correlated with other variables, using usual inferential statistical analyses common to psychology (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 (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 often 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 various 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

With language, 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 rth most common word in a given language is inversely proportional to its rank (r) in that language for a normalization constant k:

\[ P(w_r) \sim \frac{k}{r} \]  

(1)

The frequency of the rth most frequent word is roughly given by \( P(w_r) = \frac{k}{r} \), until about rank 1,000, such that the most common word (in English: “the”) has a probability of occurrence \( P(w_1) = 0.10 \) (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 remained 9,570 unique words across 258 million word instances. However, the 96 most frequent words accounted for more than 50% of word occurrences. Notably, the most common words were 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 idiocentric 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.

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 with 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. Of these, six were specifically designed to track psychological dimensions (vs. providing a generic infrastructure for counting keywords) and have more than a few hundred citations in the academic literature:

- The General Inquirer (GI; Stone et al., 1968)
- DICTION (Hart, 1984)
- Linguistic Inquiry and Word Count, 1993, 2001, 2007, 2015 (LIWC; Francis & Pennebaker, 1993; Pennebaker et al., 2007; Pennebaker et al., 2015; Pennebaker et al., 2001)
- Regressive Imagery Dictionary/Count (Martindale, 1973)
- TAS/C (Mergenthaler & Bucci, 1999)
- Gottschalk-Gleser Scales (Gleser et al., 1961; Gottschalk & 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 (2007: Pennebaker et al., 2007; 2015: Pennebaker et al., 2015; 2001: Pennebaker et al., 2001) were cited 8,800 times. The primary citations for GI (Stone et al., 1962; Stone et al., 1968) 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 multipurpose 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., 1968). Between 1962 and 1965, over 25 dictionaries were developed, with additional dictionaries developed over subsequent decades. The latest version includes 182 dictionaries (see online supplemental materials for a full list of 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 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 included within 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., 1968). 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 et al., 1964).

**DICTION**

DICTION was developed in the 1980s to analyze the “verbal tone” in 500 U.S. 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 nonoverlapping 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 online supplemental materials 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 with these norms. Importantly, DICTION was 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. For instance, 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 comprise sadness, anxiety, and anger dictionaries. As a result, when subdictionaries (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 rests 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 as the increased use of first person singular “I” pronouns tracking lower status in dyadic interactions (e.g., Campbell & Pennebaker, 2003; Chung & 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).

² When determining the number of words contained within a set of dictionaries, we counted relevant word stems (e.g., for happy, we included “happy”, “happier”, and “happiness”). Words can appear in multiple dictionaries.
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 nonzero). These make them better suited as features in predictive models across a variety of tasks in natural language processing (NLP) and sometimes provide interpretable abstractions of language in the form of word groups (or topics).

LSA and 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 et al., 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, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013).

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 multidimensional space whereby 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 is 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 microdictionaries 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 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.

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-occurrence 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. 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”)

3 For example, see http://wwwb.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.
seek to discover the different contexts in which words 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 whether or not a word occurs). Vector semantic approaches are fundamentally based on the distributional hypothesis that states: “words that occur in similar contexts tend to have similar meanings” (Jurafsky & Martin, 2020, p. 1).

LSA employed dimensionality reduction to a global word-by-document matrix, such that each row 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 (such as Word2Vec, Mikolov et al., 2013; and GloVe, Pennington et al., 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 three to six words on either side of the target word being embedded (Jurafsky & Martin, 2020; Mikolov et al., 2013). 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 datasets (even larger than those for LDA because the computational processing is less intensive), and then become fixed vector representations that can be extracted from smaller study datasets. This has been the key to the success of these approaches—they have been pretrained 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 distinctive language contexts by groups with access to

---

4 This general idea of trying to predict missing words, so-called “self-supervised learning,” remains dominant in how the state-of-the-art word embeddings are trained, even as the statistical models that are used have evolved considerably (e.g., BERT; Devlin et al., 2019).
the largest computational resources, such as Google Research (e.g., Devlin et al., 2019; Pennington et al., 2014).

Similar to LSA, the distance between the vectors of two words in the embedding space captures semantic similarity of those words. In a psychological application, Bhatia (2017) demonstrated that these semantic distances predict the association between concepts observed across a variety of judgment tasks. Specifically, the semantic distances appear to capture the associations that 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 et al. (2017) clustered the semantic distances between the vector representations of adjectives describing situations to find support for a seven-dimensional taxonomy of situations. Thus, it appears that embeddings recover regularities in our mental and physical worlds that are encoded in natural language.

The embedding vectors have also proven 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 (e.g., support vector machines; random forests; ridge regression) or deep learning systems. As an example, the differences (“offsets”) between 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, 2020). Word embeddings (and now contextual word embeddings) have become the de facto input for most NLP 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 that the word could play, without knowing the exact context in which it is being applied. It will undoubtedly contain information irrelevant to the current context (e.g., the word “bank” might capture the idea of a financial institution but is used in the sentence, “The river rose high on the bank”). However, a new generation of embeddings, contextual word embeddings, produce vectors that are specific to the context in which the word is being applied. For example, fixed embeddings assign the same vector to “play” for both “they played soccer” and “they went to the play.” With contextual word embeddings, once they are learned (pretrained) on giga-byte-scale datasets, they can assign a different embedding to each instance of “play” which better captures its sense, based on the context. Thus, unlike fixed word embeddings, contextual word embeddings require context to be considered during extraction, not just during learning, and thus are computationally more intensive. While smaller scale versions of contextual embeddings have existed for decades (e.g., Dhillon et al., 2011; Leacock et al., 1993; Schwartz & Gomez, 2008), the recent wave of contextual embeddings are based on highly complex deep learning models such as bidirectional multilayer recurrent neural networks (ELMO; Peters et al., 2018) or 12-layer transformer networks (BERT, Devlin et al., 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 (Devlin et al., 2019; Peters et al., 2018). As of 2020, word embeddings have only rarely been used in the psychological literature (e.g., Bhatia, 2017; Bhatia et al., 2019; Kern et al., 2019; Richie et al., 2019) and contextual embeddings have yet to be published in the top general psychology journals such as Psychological Science or Journal of Experimental Psychology: General.

Differential Language Analysis

LSA, LDA, and the various embedding methods cluster language into lower dimensional representations of features. DFA, on the other hand, is a relatively simple method that explores the associations of language features with extralinguistic author or text attributes of interest, such as personality traits. As such, it can use language clusters as features, or individual words and multiclass 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, DFA is sensitive toward emoticons (e.g., :-), emojis and punctuations (e.g., !!!), and misspellings, which is important for use with social media. It also includes multiclass expressions (n-grams or phrases), or a set of words that commonly occur together (e.g., “happy new year”). (For a full overview of the method, see Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013. For examples of DFA applied to personality, age, and gender, see Kern, Eichstaedt, Schwartz, Dziurzynski, et al., 2014; Kern, Eichstaedt, Schwartz, Park, et al., 2014, 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). DFA 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 DFA fundamentally is 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

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. DFA better captures dynamic changes and idiosyncrasies of online language use.
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 may thus be 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 identify psychologically relevant differences 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. Given that both approaches have strengths and weaknesses, 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 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 et al. (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 GL, 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; Ilić 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, Eichstaedt, Kern, Dziurzynski, Ramones, et al. (2013) 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 et al. (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’ behaviors 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 traits are 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 et al., 2013).

Method

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

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.

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 to allow 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.
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 Big Five personality traits, as well as their capacity to quantitatively capture variance in these variables. 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 nonconventional usages and spellings (e.g., “omg”, “wtf”), punctuation, and emoticons (e.g., :-) , ^_^), 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-gram) and three (3-gram) 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. For example, the phrase “happy birthday” was much more likely than the independent probabilities of “happy” and “birthday”. We used the pointwise mutual information (PMI) criterion to quantify these probabilities, keeping phrases with a threshold above three (for a full discussion, see Kern et al., 2016 and Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013). 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.,
etc., emojis). We used the simpler, program-based ex-
traction method for our correlational 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 online supplemental 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 in-
cluded multiple word endings as dictated by the dictionaries
(e.g., *happ* included “happy, happier”, and “happiness”).

**Topic Extraction**

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

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

\[
p(\text{topic} | \text{user}) = \sum_{\text{words} \in \text{topics}} p(\text{topic} | \text{word}) \cdot p(\text{word} | \text{user})
\]  

(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 summa-
ize results. Regression analyses compared predictive validity. We
also considered necessary samples sizes and the utility of extract-
ing 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 Big Five personality traits 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 explor-
atory and any “significant” values could be due to chance. Given
the large number of regressions, we corrected for multiple com-
parisons using the Benjamini-Hochberg procedure (BH; Benjamini
& Hochberg, 1995), which corrects the customary significance
threshold \( (p = .05) \) for the number of features that are simultane-
nously 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 underestimating 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, the size of the word indicates the frequency of
occurrence, and color is meaningless. We used DLATK to gener-
ate modified word clouds that scale the words by the magnitude of
their correlation coefficient, such that larger words indicate stron-
ger correlations with the outcome, and color indicates frequency,
from red (frequently used) to blue (moderately used) to gray
(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, Eichstaedt, Blanco, et al.,
2013).

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 preva-
ience (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, Schwartz,
Eichstaedt, et al., 2015; Sap et al., 2014; Schwartz, Eichstaedt,
Kern, Dziurzynski, Ramones, et al., 2013), 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; Hoel & Kennard, 1970) for the continuous age
and personality variables. Both techniques are straightforward

---

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.
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 intercorrelations 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 online supplemental materials 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).

Language Profiles of Gender, Age, and Personality

Figure 4 provides a quantitative summary of the correlates of gender, age, and Big Five personality traits across the five methods. The figure provides the 10 largest positive and negative standardized regression coefficients between the dictionaries and outcomes 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 (e.g., “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 (e.g., “battle”, “victory”, “fight”), political debate (e.g., “country”, “power”, “freedom”), and sports (e.g., “football”, “season”, “team”; “win”, “lose”, “bet”).

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 subdictionaries as well. If the broader dictionary did not show a significant association but two or more subdictionaries were significant, we note the higher order dictionary but leave the coefficient blank.
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 (e.g., “wonderful”, “amazing”, “thank you”), social connection and events (e.g., “friends”, “family”, “weekend”, “thanksgiving”), and religiosity. The language of disagreeableness included cursing and negative appraisals of others (e.g., “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 included references to family and friends (e.g., “family”,

---

### Table 1

**Intercorrelations Among Positive Affect, Negative Affect, and Pronoun Dictionaries**

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>General Inquirer</th>
<th>LIWC 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lasswell</td>
<td>Harvard IV</td>
</tr>
<tr>
<td></td>
<td>Positive affect</td>
<td>Pleasure</td>
</tr>
<tr>
<td>General Inquirer</td>
<td>.48</td>
<td>.70</td>
</tr>
<tr>
<td>Pleasure</td>
<td>.37</td>
<td>.47</td>
</tr>
<tr>
<td>Positive</td>
<td>.45</td>
<td>.60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>General Inquirer</th>
<th>LIWC 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lasswell</td>
<td>Harvard IV</td>
</tr>
<tr>
<td></td>
<td>Negative affect</td>
<td>Vice</td>
</tr>
<tr>
<td>General Inquirer</td>
<td>.59</td>
<td>.68</td>
</tr>
<tr>
<td>Negative</td>
<td>.38</td>
<td>.26</td>
</tr>
<tr>
<td>Hostile</td>
<td>.30</td>
<td>.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>General Inquirer</th>
<th>LIWC 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Diction</td>
<td>LIWC 2015</td>
</tr>
<tr>
<td></td>
<td>Harvard IV: Self</td>
<td>Self-reference</td>
</tr>
<tr>
<td>Diction</td>
<td>Self-reference</td>
<td>.75</td>
</tr>
<tr>
<td>LIWC</td>
<td>Pronouns</td>
<td>.75</td>
</tr>
<tr>
<td>Personal Pronouns</td>
<td>.70</td>
<td>.60</td>
</tr>
<tr>
<td>First person singular</td>
<td>.92</td>
<td>.80</td>
</tr>
</tbody>
</table>

**Note.** DICTION and LIWC2015 dictionaries were extracted through their respective programs, GI dictionaries were extracted through DLATK. LIWC = Linguistic Inquiry Word Count program; GI = General Inquirer.
“friends”, “blessed”), structured social time (e.g., “weekend”, “spending”, “hanging”), and relaxing from work (e.g., “relaxation”, “vacation”, “recover”). 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 (e.g., “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 (e.g., “feeling”, “tired”, “sick”), hostility and cursing, exhaustion and overarousal (e.g., “stressed”, “frustrated”, “annoyed”), and depressed mood. Emotional stability (low neuroticism) was distinguished by mentions of weekends (e.g., “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 (e.g., “human”, “nature”, “universe”, “wonders”) and artistic (e.g., “writing”, “write”, “poetry”) concerns. Low openness was associated with pragmatic, domestic concerns including home, family, and temporal concepts.

**LIWC2007 Versus LIWC2015**

As noted above, 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 online 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.

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 on the study dataset (Park, Schwartz, Eichstaedt, et al., 2015; 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 dataset (Lynn et al., 2020). Finally, we include Azucar et al.’s (2018) meta-analytic estimates for prediction accuracies for social media-based prediction of Big Five personality traits across datasets.

As shown in Table 2, DICTION’s dictionaries captured less information about personality ($r_{\text{average}} = .23$) than the LIWC ($r_{\text{average}} = .28$) and GI ($r_{\text{average}} = .29$) dictionaries. 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 ($r_{\text{average}} = .37$ to 39) and nominally above the meta-analytic baseline ($r_{\text{average}} = .35$). This suggests that all of these approaches capture a similar amount of language variance, but that the word embeddings are more parsimoniously, with fewer language dimensions.

<table>
<thead>
<tr>
<th>LDA Topics Most Associated with Older Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA Topics Most Associated with Younger Age</td>
</tr>
</tbody>
</table>

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

Figure 4. (continued)
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 online supplemental materials 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 amongst the traits; for example, for openness, 500 users sufficed for 100 significantly associated LDA topics, whereas for neuroticism, a sample of 1,800 was needed.

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 online supplemental materials for each personality dimension, and for age and gender). Generally, sample size and number of words per user were a 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 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.

Note. All coefficients are significant at p < .001, corrected for multiple comparisons. (m) designates “master” categories that combine frequencies of multiple dictionaries.
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, Eichstaedt, Blanco, et al., 2013), 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 online supplemental 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 online supplemental 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. We recommend that users also manually check the most frequent words within the dictionaries being used (see online supplemental materials),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 5 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.

Discussion

There is a raft of remarkable work in language analysis in fields related to psychology, including style matching (Gonzales et al., 2010; Ireland et al., 2011; Taylor & Thomas, 2008); how power differentials among 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 et al., 2011; Strapparava & Mihalcea, 2008) and psychological traits (Guntuku et al., 2017; Mitchell et al., 2015; Park, Schwartz, Eichstaedt, 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 et al. (2003) foresaw, “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 (see for instance: Box, 1976; Kennedy-Shaffer, 2019). Similarly, closed-vocabulary approaches originated at a time when very few large-scale correlational studies existed, limited amounts of text were recorded in experimental contexts, and qualitative information was hard to capture (Anderson et al., 2012). 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.

---

**Table:**

<table>
<thead>
<tr>
<th>LDA Topics Most Associated with High Neuroticism</th>
<th>LDA Topics Most Associated with Low Neuroticism</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta = .90 )</td>
<td>( \beta = .07 )</td>
</tr>
<tr>
<td>( \beta = .08 )</td>
<td>( \beta = .08 )</td>
</tr>
<tr>
<td>( \beta = .08 )</td>
<td>( \beta = .08 )</td>
</tr>
<tr>
<td>( \beta = .08 )</td>
<td>( \beta = .07 )</td>
</tr>
</tbody>
</table>

---

**Note.** All coefficients are significant at \( p < .001 \), corrected for multiple comparisons. (m) designates “master” categories that combine frequencies of multiple dictionaries.
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, Eichstaedt, Schwartz, Park, et al., 2014; Pennebaker & Stone, 2003; Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013), gender (e.g., Newman et al., 2008; Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013), and personality (Kern, Eichstaedt, Schwartz, Dziurzynski, et al., 2014; Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013; Yarkoni, 2010). GI, DICT-ION, 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 et al., 2009; Golbeck et al., 2011; Sumner et al., 2011). Similar to previous work (Iacobelli et al., 2011; Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013), the open-vocabulary prediction models outperformed dictionary-based prediction models, suggesting that the larger number of open-vocabulary features capture more of the

---

**Table: LDA Topics Most Associated with High Openness**

<table>
<thead>
<tr>
<th>Topic</th>
<th>β</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change</td>
<td>.16</td>
</tr>
<tr>
<td>Clout (m)</td>
<td>.13</td>
</tr>
<tr>
<td>Consequences</td>
<td>.13</td>
</tr>
<tr>
<td>Emotion</td>
<td>.12</td>
</tr>
<tr>
<td>Emotion</td>
<td>.12</td>
</tr>
</tbody>
</table>

**Table: LDA Topics Most Associated with Low Openness**

<table>
<thead>
<tr>
<th>Topic</th>
<th>β</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggression</td>
<td>.11</td>
</tr>
<tr>
<td>Achievement</td>
<td>.11</td>
</tr>
<tr>
<td>Affiliation</td>
<td>.11</td>
</tr>
<tr>
<td>Agreement</td>
<td>.11</td>
</tr>
<tr>
<td>Agreement</td>
<td>.10</td>
</tr>
</tbody>
</table>

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

---

**Note.** (Figure continues on next page.)
The use of dictionaries in text analysis has increased in recent years, with researchers using them for a wide range of purposes. These approaches can be divided into two main categories: closed-vocabulary and open-vocabulary. Closed-vocabulary approaches use a predefined set of words, phrases, or topics to analyze text, while open-vocabulary approaches use a larger, more flexible set of words or phrases.

Choosing an Approach

Closed-vocabulary programs are often used in psychological research, as they can capture the nuances of everyday language processes. This is fundamentally different from the closed-vocabulary approaches that were initially used for analyzing reflective essays or presidential speeches. The purpose of the research (for which DICTION was created) is a driving force in selecting the approach.

Open-vocabulary approaches yield more specific language inferences than closed-vocabulary approaches. They are particularly suited for capturing 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 fundamental 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.


table

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Diction</th>
<th>LIWC 2015</th>
<th>General Inquirer</th>
<th>LDA Topics</th>
<th>LDA Topics, Words, Phrases</th>
<th>Word2Vec Embeddings</th>
<th>BERT Embeddings</th>
<th>Meta-analytic estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of language vars.</td>
<td>31</td>
<td>73</td>
<td>182</td>
<td>2,000</td>
<td>&gt;10,000</td>
<td>200</td>
<td>768</td>
<td>(various studies)</td>
</tr>
<tr>
<td>Age (r)</td>
<td>.56 (.55, .56)</td>
<td>.65 (.65, .66)</td>
<td>.68 (.68, .69)</td>
<td>.81 (.81, .81)</td>
<td>.83 (a)</td>
<td>.92 (a)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender (accuracy)</td>
<td>.00 (.74, .75)</td>
<td>.78 (.78, .79)</td>
<td>.82 (.81, .82)</td>
<td>.89 (.89, .89)</td>
<td>.92 (a)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agreeableness (r)</td>
<td>.21 (.20, .22)</td>
<td>.26 (.25, .27)</td>
<td>.25 (.24, .26)</td>
<td>.32 (.32, .33)</td>
<td>.35 (b)</td>
<td>.33 (c)</td>
<td>.37 (c)</td>
<td>.29 (.21, .36)</td>
</tr>
<tr>
<td>Conscientiousness (r)</td>
<td>.26 (.26, .27)</td>
<td>.28 (.27, .28)</td>
<td>.31 (.30, .31)</td>
<td>.37 (.36, .37)</td>
<td>.37 (b)</td>
<td>.37 (b)</td>
<td>.38 (b)</td>
<td>.35 (.29, .42)</td>
</tr>
<tr>
<td>Extraversion (r)</td>
<td>.22 (.21, .23)</td>
<td>.30 (.29, .31)</td>
<td>.30 (.29, .30)</td>
<td>.38 (.38, .39)</td>
<td>.42 (a)</td>
<td>.37 (c)</td>
<td>.39 (c)</td>
<td>.40 (.33, .46)</td>
</tr>
<tr>
<td>Neuroticism (r)</td>
<td>.20 (.19, .21)</td>
<td>.24 (.23, .25)</td>
<td>.27 (.26, .27)</td>
<td>.34 (.33, .35)</td>
<td>.35 (b)</td>
<td>.37 (c)</td>
<td>.38 (c)</td>
<td>.35 (.27, .39)</td>
</tr>
<tr>
<td>Openness (r)</td>
<td>.26 (.25, .26)</td>
<td>.30 (.30, .31)</td>
<td>.33 (.32, .33)</td>
<td>.43 (.43, .44)</td>
<td>.43 (b)</td>
<td>.39 (c)</td>
<td>.44 (c)</td>
<td>.39 (.30, .48)</td>
</tr>
<tr>
<td>Average personality (r)</td>
<td>.23</td>
<td>.28</td>
<td>.29</td>
<td>.37</td>
<td>.38</td>
<td>.37</td>
<td>.39</td>
<td>.35</td>
</tr>
<tr>
<td>Average pers. adj. R²</td>
<td>.05</td>
<td>.08</td>
<td>.08</td>
<td>.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 ³ Sap et al. (2014) and ⁴ Park, Schwartz, Eichstaedt, et al. (2015); for vector semantic (Word2Vec) and contextual (BERT) embeddings in ⁵ Lynn et al. (2020) disattenuated for measurement reliability (r = .734). For BERT, we report the BERT + DAN model. Meta-analytic estimates are based on those reported in Azucar et al. (2018). Parentheses indicate 95% confidence intervals. LIWC = Linguistic Inquiry Word Count program; LDA = Latent Dirichlet Allocation; BERT = Bidirectional Encoder Representations from Transformers.

One advantage of dictionary-based methods is their relatively small number of language features (i.e., the number of selected dictionaries), compared with the very large number of words, phrases, and topics that occur with 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 to understand mechanisms and test a priori theories, 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 demographic and personality variables, language profiles that capture many of the specific distinctions among users were observed with similar sample sizes for the LDA topics and the LIWC2015 dictionaries (N/H11011250 vs. 200 for demographics, N/H110111,000 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/H11011650 for demographics and N/H110113,000 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, inversely, if more words per user are available, smaller sample sizes may be adequate. For example, distinctive language profiles for age and gender for LIWC and LDA topics could be observed with as few as 20 – 40 words per user for a sample of N/H110055,000 users, while for a sample of N/H11005150, thousands of words per user were required (see Table 4). Generally, more textual data is required to explore the language of personality than for demographics. As a rule of thumb, for personality traits, for both LIWC and topics, for a sample of N/H110051,000, thousands of words are needed from a user. For a sample of N/H110055,000, hundreds of words may suffice. As

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

Note. 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. See the online article for the color version of this figure.
reported above, to identify meaningfully distinctive 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 dataset is about 21 words long, and an average Tweet is about 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 online supplemental materials [https://osf.io/h4y56] for more detailed figures about when 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 noncommercial 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

**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 |
|-------------------------------------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| Demographics | Big Five personality traits |
| thresholds of significant correlates | Gender | Age | Agreeableness | Conscientiousness | Extraversion | Neuroticism | Openness (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). LIWC = Linguistic Inquiry Word Count; LDA = Latent Dirichlet Allocation.

**Figure 6**

Average Number of Language Features Significantly Associated Across Personality Dimensions as a Function of Words per User (Log-Transformed)

*Note.* 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 online supplemental materials for additional figures. See the online article for the color version of this figure.
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, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013; 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 modeled, 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 datasets with millions of documents, we have found that 500–2,000 topics provide the right level of distinctive detail, with the most correlated topics visualized to yield a general view of what users are writing about. Larger numbers of topics (in the thousands) will

<table>
<thead>
<tr>
<th>Sample sizes</th>
<th>N = 150</th>
<th>N = 1,000</th>
<th>N = 5,000</th>
<th>N = 150</th>
<th>N = 1,000</th>
<th>N = 5,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 (out of 73) LIWC dictionaries</td>
<td>~4,000+</td>
<td>90+</td>
<td>20+</td>
<td>—</td>
<td>—</td>
<td>1,000 to 4,000+</td>
</tr>
<tr>
<td>100 (out of 2,000) LDA topics</td>
<td>~4,000+</td>
<td>150+</td>
<td>40+</td>
<td>—</td>
<td>—</td>
<td>~4,000+</td>
</tr>
<tr>
<td>200 (out of 11,894) 1-to-3 grams</td>
<td>—</td>
<td>750+</td>
<td>240+</td>
<td>—</td>
<td>—</td>
<td>1,000 to 4,000+</td>
</tr>
</tbody>
</table>

*Note.* For missing values, the threshold number of meaningful associations was not reached even when including all of the users’ language (an average of 4,104 words). LIWC = Linguistic Inquiry Word County program; LDA = Latent Dirichlet Allocation.
This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

While many packages exist to perform topic modeling (such as Mallet; McCallum, 2002), none of them currently are as easy 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). DLATK can also be used for psychological insight. In this review, we have discussed computational linguistics in psychology are fairly new, these 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 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., multilevel LDA, or a combination of LDA topic modeling with matrix factorization 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 could be on the horizon.

Word and contextual embedding models are just beginning to be used for psychological insight. In this review, we have discussed another way to automatically determine the number of topics to be modeled (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 are 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). DLATK can also be carried out online (http://lexhub.org). In addition, in the online supplemental 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 much trait-related variance as LIWC, but are free for noncommercial 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, Eichstaedt, Blanco, et al., 2013). 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., multilevel LDA, or a combination of LDA topic modeling with matrix factorization 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 could be on the horizon.

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 are 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). DLATK can also be carried out online (http://lexhub.org). In addition, in the online supplemental 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 much trait-related variance as LIWC, but are free for noncommercial 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, Eichstaedt, Blanco, et al., 2013). 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., multilevel LDA, or a combination of LDA topic modeling with matrix factorization 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 could be on the horizon.

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 are 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). DLATK can also be carried out online (http://lexhub.org). In addition, in the online supplemental 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 much trait-related variance as LIWC, but are free for noncommercial 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, Eichstaedt, Blanco, et al., 2013). 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., multilevel LDA, or a combination of LDA topic modeling with matrix factorization 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 could be on the horizon.

Word and contextual embedding models are just beginning to be used for psychological insight. In this review, we have discussed

---

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

<table>
<thead>
<tr>
<th>Topic set</th>
<th>Occasions</th>
<th>Top 10 words comprising each topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>1</td>
<td>game, play, win, playing, football, team, won, games, beat, lets</td>
</tr>
<tr>
<td>500</td>
<td>5</td>
<td>guitar, play, playing, music, piano, band, bass, hero, practice, played</td>
</tr>
<tr>
<td>2,000</td>
<td>9</td>
<td>play, guitar, learn, piano, learning, playing, learned, lessons, piano, rules</td>
</tr>
</tbody>
</table>

Note. Words suggesting playing music are highlighted in green, ball sports in blue, and videogames in yellow. See the online article for the color version of this table.
DLA, 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. 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.

Throughout this article, we have observed that off-the-shelf dictionaries may often be suitable to test specific hypotheses. However, in situations where such training data are 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 available (e.g., temporal orientation: Park, Schwartz, Sap, et al., 2015, valence/arousal: Eichstaedt & Weidman, 2020, or empathy: Abdul-Mageed et al., 2017). The evolution 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, where 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 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.

---

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


Received March 5, 2019
Revision received July 7, 2020
Accepted July 13, 2020

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.