7

TEXTUAL ANALYSIS

Cindy K. Chung and James W. Pennebaker

Introduction

Language is a defining feature of human culture. Although social scientists have long agreed about the profound nature of language, they have been reticent to study it. And there is good reason: it is exceedingly difficult to record, amass, extract, and to code or decode what people are saying. Only recently have social scientists made giant strides in measuring and analyzing the words people use in their everyday lives. Twenty-first-century technological advances have given birth to computer-mediated communication (CMC), which has made our lives considerably easier. For the first time in history, we have the tools to track human interaction through the spoken and written word quickly and efficiently, and at a scale that was unimaginable even a decade ago.

CMC is providing social psychologists with new questions to answer about micro-interactions, emotional tone, group dynamics, and cultural shifts that may change in seconds or centuries. The analysis of CMC and other types of language offers a means to understand how we are influenced by the actual, implied, or imagined presence of others. We can now analyze millions of books and manuscripts dating back centuries. We can also quickly track societal changes in thinking and communication through the analysis of language in Snapchat, Facebook, Tinder, or one of hundreds of new apps that appear each year both within and across cultures.

Even the term "CMC" seems rather antiquated. Most communication now occurs over some type of digitally connected device, with hand-written letters considered a dying art, an unscheduled phone call too obtrusive, and commitments that don't really exist unless by email or SMS confirmation. The majority of humans in developed countries own a personal smartphone for text messaging, manage several cluttered email inboxes with thousands of unread messages, and are similarly guilty for having an ever-growing email outbox. It is customary for individuals who have never met or spoken in person to interact long term daily and digitally. For example, these might be primary work collaborators, online acquaintances explicitly looking for love, 20 hours per week video game allies, opposing lawsuit parties, devoted customer and merchant, or daily content makers and subscribers.

Communication, and by extension, social interactions, have changed. Although the medium has changed—as have the tools to record, amass, extract, code or decode, and to assess their meaning or style—the words, especially those that are most revealing of social dynamics, have largely stayed the same.

Content vs. Function Words

A helpful way to look at measuring language in social psychology is to consider two broad categories: content words and function words (see Pennebaker, 2011). Content words are made up of nouns, regular verbs, adjectives, and adverbs. Content words tell us what people are thinking about. Function words are made up articles, auxiliary verbs, conjunctions, negations, pronouns, and prepositions. Function words tell us how people are thinking and connecting with others.

Both categories of words are revealing of our thoughts, feelings, and behaviors, with function words being more reliable markers of psychological states and social dynamics across topics (Chung & Pennebaker, 2007; Mehl, 2006; Pennebaker, Mehl, & Niederhoffer, 2003), which are typically represented by content words. For example, the topic of the statement

My Facebook post had a lot of comments.

is gleaned from the words "Facebook," "post," and "comments."

How someone is thinking about the topic is understood from "My," "had," and "a lot of." Specifically, "My" represents self-focus: a personal share or ownership of the topic. Had the speaker used "The" instead of "My," the Facebook post could've been written by anyone, including a more personally distanced way of referring to the speaker's own Facebook post. "Had" represents past-tense focus. Together, "my" and "had" assume a shared reference between the speaker and the audience as to which of the speaker's past Facebook posts the speaker is referencing.

"A lot of" represents some comparison to a quantity which is unknown unless the speaker and audience have a shared reference point to how many comments represent a relatively large quantity. Had the speaker used "a lot of" knowing that both the speaker and audience thinks that say, over 50 comments is a large quantity, but the speaker had in fact received 2 comments, this statement might be viewed as sarcastic or funny as opposed to a casual and intentionally accurate statement about the large (i.e., over 50) quantity of comments. Had the speaker used "a lot of" thinking that the audience was interested in frequent reports on the number of comments received on each of the speaker's Facebook posts, but the audience was, in fact, not interested in said reports, this statement might be viewed as annoying or gratuitously boastful.

Note that it is not very interesting to think about the meaning of each word as in the laborious example above. (Note also, that linguists may disagree.) However, a few takeaways from the exercise are that (a) speakers and listeners process function words automatically without going through the steps above. (b) Function words are inherently social, drawing on shared references between a speaker or writer and their audience to use and to understand. (c) Different psychological states and social dynamics are associated with different categories of function word use. (d) Function words make up over 50% of the words we use in our everyday speech and writing. Together, these make function words excellent observable behaviors to understand how individuals are influenced by the actual, implied, or imagined presence of others. In other words, function words are the stuff of social psychologists' dreams.

Linguistic Inquiry and Word Count (LIWC)

Admittedly, most social psychologists don't, in fact, dream about function words as gateways into the inner workings of social dynamics. It is understandable why this may be the case. As mentioned in the introduction, language has not always been easy to record, amass, extract, code or decode, and to assess its meaning or style. However, several innovations have made the analysis of function words much more accessible to social psychologists. The primary tool that turned widespread attention in social psychology to function words was the advent of Linguistic Inquiry and Word Count (LIWC 2001; Pennebaker, Francis, & Booth, 2001). LIWC, pronounced "Luke," is a software made up of a processor and a dictionary. The processor counts words in the category entries listed in the dictionary, and reports on the percentage of words in each text file that represents each dictionary category.

The standard LIWC dictionary is made up of over 80 categories, including function word categories (e.g., articles, negations, pronouns, etc.), and content word categories (e.g., positive and negative emotion words, cognitive mechanisms, social words, biological words, achievement, religion, etc.). Each of the words in the standard LIWC dictionary having been judged by four judges, and agreed on by at least three of those judges, as belonging to its category. It is possible to have the processor count words in custom dictionaries: this function makes it possible for users to create their own "dictionary," made up of words of their choosing in categories of their choosing.

LIWC is relatively easier to use than other natural language processing (NLP) techniques as LIWC is a cross-platform application, and no programming is required. We refer the reader to the LIWC website (www.liwc.net) for a manual on its use but provide a brief overview here. For any given project, a corpus (a body of text files) is collected into a folder, where each text file in the corpus represents an observation (i.e. all the typed or transcribed words of an individual, or, a single message from an individual). Each text file should have a minimum number of words specified by the researcher. Note that LIWC reports on the percentages of words, and so a minimum cutoff around 100 words may seem reasonable for many studies, although there may be reasons to decrease this cutoff, and there are no hard and fast rules on what should be the minimum cutoff. In short, more words are associated with greater reliability.

Within LIWC, the first step is to identify the location of the text files to process. The default dictionary is referenced by default, although, as previously mentioned, it is possible to have LIWC point to a custom dictionary. Once the files are selected, LIWC automatically processes all files, counting the percentage of words in each category of the LIWC dictionary for each text file.

The LIWC output, which is a matrix of text files in rows, LIWC categories in columns, and percentages of use in each cell, is saved as an output file in TXT, CSV, or Excel format. This output file can be opened with any statistical package to conduct analyses on the relative rates of word use by different groups of text files. Accordingly, it is essential to design an empirical study with statistical tests that will answer one's research questions in advance of preparing the text files, just as any survey or observational study would format data collection to have observations in rows, measurements in columns, and values in cells. The statistical analyses to be conducted are entirely dependent on the research questions posed. Ultimately, having a large collection of words is not enough; having an appropriate understanding of experimental design and a statistical analytic strategy, just as in any empirical social psychological assessment, are required.

Several updates, including the product's commercialization, have been made to both the processor (Pennebaker, Booth, & Francis, 2007) and the standard LIWC dictionary in 2007 Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007) and in 2015 (Pennebaker, Booth, Boyd, & Francis, 2015). The current 2015 processor has the ability to process text in various file formats such as .xlsx, .csv, .pdf, and others, beyond plain ASCII text files. Currently, the standard LIWC dictionary (from both earlier and the current versions) has been translated into Arabic, Chinese, Dutch, French, German, Korean, Russian, Spanish, and Turkish (see www.liwc.net), with several other languages under development. The latest standard LIWC dictionary in English includes several super categories (e.g., analytic thinking, authenticity, etc.) which are constructs derived from LIWC categories and based on past research that has used LIWC (Pennebaker, Boyd, Jordan, & Blackburn, 2015).

Note that there are a myriad of other theoretically based computerized word counting programs, including The General Inquirer for psychological topics (Stone, Dunphy, Smith, & Ogilvie, 1966), DICTION for political texts (Hart & Carrol, 2014), and TAS/C for psychotherapeutic transcripts (Mergenthaler, 1996; Mergenthaler & Bucci, 1999). There are also an increasing number of text analysis methods from the broader field of NLP that can be used to measure words. There are an overwhelming number of text analysis program to list. So, in order to avoid overwhelming a beginner to text analysis, we provide a table below on just a few of the popular, well-maintained, and simplest tools for which no programming is required. These tools can easily be used by social scientists to begin to incorporate some text analysis in their multi-methods toolkits.

Many of these text analytic software packages count and categorize words; they are measurement tools. However, to gain descriptive and inferential insight into what those words signal, descriptive and inferential statistics must be applied in a subsequent step, in a statistical software package.

TABLE 7.1 Examples of Text Analysis Software

Tool	*Purpose*	*Reference*
Linguistic Inquiry and Word Count (LIWC)	To derive a matrix of words that indicates the percentages of words in a text belonging to validated categories (grammatical, psychological, content); custom dictionaries can be uploaded to assess a corpus for specific words; comparisons in word use across and between groups can be conducted when the matrix is uploaded to a statistical software package.	www.liwc.net
Meaning Extraction Method (MEH)	To derive a matrix of terms that indicates the percentages of words in a text belonging to the most frequently referenced terms in a corpus; facilitates the Meaning Extraction Method (MEM; Chung & Pennebaker (2008); a method to inductively derive topics in a corpus) when the matrix is uploaded to a statistical software package.	http://meh.ryanb.cc
QDA Miner Provalis	To conduct frequency and percentage counts of words in a corpus that are not necessarily dictionary driven; some descriptives and basic topic modeling capabilities for descriptive purposes.	https://provalisresearch.com
WordSmith	To conduct frequency and percentage counts of words in a corpus that are not necessarily dictionary driven, but more descriptive of a text; co-occurrences of words can also be assessed.	http://lexically.net
Coh-Metrix	To assess over a hundred features of text using a variety of statistically derived indices, including cohesion, and readability.	www.cohmetrix.com

In the case of LIWC, recall that one of the requirements for the inclusion of words in the LIWC dictionary was that judges agreed that a word belonged in a category. Since function words are widely agreed upon fixed lists, these were added to the standard LIWC dictionary. It wasn't until after using LIWC for a wide variety of studies that function words were often found to be more reliable markers of psychological state than content words (Pennebaker, 2011). With word counting as a basic foundation of quantitative text analysis across many disciplines (see O'Connor, Banman, & Smith, 2011), a growing complexity of statistical techniques have been applied to word counts including the output of LIWC's categories to derive new insights into human behaviors.

Note that there is software to crawl the web, process text, and compute complex statistical procedures on the text, such as TACIT (e.g., Deghani et al., 2016). However, for the text analysis beginner, starting with simple word counts, and applying familiar statistical techniques typically used in social psychological studies might ease the understanding of text analysis as a tool in one's larger social science toolkit.

In the next section, we provide an overview of the studies that have used LIWC or other text analysis approaches to examine social psychological research questions. Then, we review the growth of text analysis across disciplines. Finally, we touch on some of the issues that the future of text analysis applications in social psychology faces, and future directions.

Social Psychological Applications

LIWC and other quantitative text analytic tools and strategies have been applied across a variety of topic areas within social psychology, including the assessment of status, romantic relationships, health, persuasion, forensics, and culture. As with any other measure in psychology, language should be assessed for its reliability and validity as a measure for any given construct. In addition, as with any other measure in psychology, one must consider the degree to which language is studied under naturalistic conditions, is appropriately powered, and has been investigated with multiple methods. The resources—including platforms, tools, applications, and statistical techniques—to study language are growing. Hand in hand with these technical resources, and our theoretical knowledge and empirical literature on human behaviors, social psychologists are able to make more reliable, generalizable, or nuanced statements, as well as new insights about individuals, groups, and culture.

Below, we provide an overview of select areas in which quantitative text analyses have been applied to social psychological research questions. For a review of quantitative text analysis of personality research questions, please see Chung and Pennebaker (forthcoming); Ireland and Mehl (2014).) Although we draw on research using other quantitative text analysis tools, we place a strong focus on the applications of LIWC, because it is the tool with which we are most familiar, and because it is the most widely applied quantitative text analysis tool in the social sciences.

Status

It's interesting to watch two strangers interacting from a distance. Even though one may not be able to hear what they are saying, it is possible to get a sense of their emotional states, how well they know one another, and which one is more in control of the relationship. Interestingly, the analysis of function words can reveal some of these same dynamics. Whatever the context, people's thoughts, feelings, and behaviors tend to systematically change in response to different situations. With quantitative text analysis, we can find clues to their thoughts, feelings, and behaviors about each other in the language they use.

These discoveries have been facilitated by records of language between interactants in real-time (e.g., text messages, social media posts and comments, closed captioning and advances in transcriptions, etc.). Language clues are apparent in both the content of speech but also in pronouns and other function words. Specifically, the topic of conversation may reflect the type of relationship one has with another person. For example, words relating to work collaborations may include "analysis," "deadlines," "document," "funding," "presentation," "report," and "review." These content words are likely to appear in professional discussions and relatively unlikely to bubble up in a heated romantic encounter.

How interactants are speaking with one another via function words provides a different view into relationships that can be relatively independent from the topic. For example, higher status individuals tend to use more "we," while lower status individuals use more "I" (Kacewicz, Pennebaker, Davis, Jeon, & Graesser, 2013). Using LIWC, this status differential has been observed across various types of relationships, including in military documents (Hancock et al., 2010), terrorist group member statements and interviews (Pennebaker & Chung, 2008), U.S. President Richard Nixon's Watergate tapes (Chung & Pennebaker, 2007), quarterly earnings transcripts of incoming and outgoing CEOs (Kacewicz, 2013), and even between IMs of randomly paired, unacquainted college students asked to talk or message one another (Kacewicz et al., 2013).

Kacewicz et al. found reliable language markers of status across five studies. They also found reliable effects for higher word count and more second person pronouns by higher status interactants. Effect sizes for language as markers of personality traits such as the Big Five and the Dark Triad typically tend to be lower (for a review of language markers of personality, please see Chung & Pennebaker, forthcoming; Ireland & Mehl, 2014), while effects for demographics such as age, gender, and status tend to be much stronger (for a review, see Tausczik & Pennebaker, 2010). Sir Francis Galton hypothesized why this might be so in his Lexical Hypothesis of Personality (1884).

The Lexical Hypothesis

According to the Lexical Hypothesis of Personality (see also Goldberg, 1993):

> Postulate 1: Traits that are important to our lives will be encoded in language.
> Postulate 2: The most important traits are likely to be represented as a single word in language.

If we were to extend Galton's Hypothesis beyond personality to social dynamics, we might expect a Lexical Hypothesis of Social Life:

> Postulate 1: Dynamics that are important to our social lives will be encoded in language
> Postulate 2: The most important dynamics to attend to in our social lives are likely to be represented as a single word in language.
> Postulate 3: The most important dynamics to attend to in our social lives are likely to be represented as the shortest, quickest-to-utter words in language.

The last postulate refers to function words, which tend to have a shorter number of letters than most words in our vernacular. In an interaction, function words quickly distinguish whom we should treat as the holder of power, resources, and tribal knowledge; whom we should attract as potential mates or hunting buddies; how far we should pitch our camp from them; how fast we should run from them if necessary. Even sighs and fillers, which aren't typically considered as full words in conversations, but transcribed in a few letters, can be indicative of well-being (Robbins et al., 2011) and demographics (Laserna, Seih, & Pennebaker, 2014).

Luckily, for many of us living thousands of years from the inception of formal language, function words are processed automatically in our frontal lobes, and so are read and spoken automatically. It is possible to infer these relationship attributes from the ways that people speak or write to each other, even when we ourselves are not a part of the conversation. If it's not obvious upon regular human observation, fear not, there are computerized text analysis tools such as LIWC to help decode relationships by examining pronoun use.

Relationship Dynamics

Given the intimate links between function words and social behaviors, it is not surprising that some of the most powerful and mysterious social psychological phenomena can be studied by looking at the ways people talk. In recent years, the computerized analysis of language has revealed new insights into our thinking about romantic attraction, persuasion, and emotional contagion.

We all have an intuitive sense when an interaction goes well or "clicks." We feel that we understand the other person and can practically finish each other's sentences. These close connections are sometimes common among old friendships and, on other occasions, appear out of nowhere between two strangers. In recent years, several studies have analyzed the language of a wide range of social interactions and have identified the quality of people's relationships and, as mentioned above, their relative status. Some studies have examined how the words in a speed-dating interaction may be predictive of going out on a subsequent date (Ireland et al., 2011; Ranganath, Jurafsky, & McFarland, 2009), or staying in a relationship (Slatcher & Pennebaker, 2006).

A measure of how two people are using function word categories at the same rates, or are mirroring one another in their non-conscious word use is more predictive of mutual attraction than is a measure of how two people use content word categories at the same rates (Ireland & Pennebaker, 2010). This measure has been termed language style matching (LSM). Higher LSM and greater positive emotion word use in relationships are seen in longer lasting relationships (Slatcher & Pennebaker, 2006).

What is fascinating about this simple measure of LSM is that it is not only telling of relationship longevity, but it signals coordination on a much larger scale. LSM has been found to be higher in Wikipedia discussions for articles that have higher ratings (Pennebaker, 2011). That is, Wikipedia articles were judged to be better if editors communicated similarly. The degree to which community members use function words similarly has also been found to be higher in Craigslist ads for mid-sized cities with a gini coefficient that indicates that wealth is more evenly distributed (Pennebaker, 2011). These studies suggest that function word analyses or LSM can be used as a remote sensor of a dyad or group's internal dynamics.

Persuasion

LSM also plays an important role in the social dynamics of persuasion. Matching with an opponent's language style in a political debate has been shown to influence viewers who are watching the debate. In a study of U.S. presidential debates, Romero et al. (2015) found that candidates who matched to the style of their opponent fared better in the election polls, presumably because it demonstrates perspective taking and greater fluency. These are particularly interesting effects since previous research has shown that lower status interactants match more to their higher status interactants (Danescu-Niculescu-Mizil, Lee, Pang, & Kleinberg, 2012).

Romero et al. (2015) replicated the presidential debate findings in a study of business student negotiations. Those who matched to their interactant were seen by third-party observers as having performed better or won the negotiation, and were more likely to be picked to negotiate for the third party observer. LSM, then, not only influences the dynamics and outcomes of an interaction, but it also affects the perception of an interaction by those who are mere observers.

Another study on persuasion and function words examined the Reddit forum "Change My View," where users post a position statement on any topic, provide supporting reasons, and then receive counterarguments from other users. Tan, Niculae, Danescu-Niculescu-Mizil, and Lee (2016) found that pronouns held predictive power for malleable positions over specific topics (e.g., food, government, etc.). Specifically, individuals who were more likely to change their minds used more first person singular pronouns in their posts; individuals were less likely to change their minds used more first person plural pronouns in their posts. These suggest that it was easier to change a viewpoint that was presented as being held by the self as opposed to a viewpoint that was presented as being held by many. Note, that some content words indicating power and success (e.g., completion, smile) were predictive of resistant posts, but topics (e.g., food, government, etc.) did not add predictive power to which posts led to changing the user's mind.

Another way in which influence has been studied using LIWC has been through the examination of word use on millions of posts on Facebook News Feeds (Kramer, Guillory, & Hancock, 2014). Specifically, an experimental study that systematically reduced the presentation of posts with LIWC's positive and negative emotion word categories showed that emotional contagion propagates via words in the absence of non-verbal cues. That is, a reduction in the proportion of posts seen with positive or negative emotion word use led to a significant reduction in corresponding positive and negative emotion word expression respectively by connections that were exposed to the experimental manipulations relative to controls. In addition, there was a significant increase in expressing the opposite emotion by connections that were exposed to the experimental manipulations relative to controls.

While the effect sizes were small, the fact that social networks are by definition interconnected suggests that the effects can be wide reaching. The words we use can have profound effects on how the people around us experience their worlds, and in turn, how they influence those around them. This is increasingly important to attend to as we increase our interactions over CMC, and in increasingly connected media.

Forensics

Language markers for forensic analyses have been identified and applied to open-ended statements, emails, papers, and conversations. For example, in studies where participants have been asked to lie in laboratory studies (e.g., Hancock, Curry, Goorha, & Woodworth, 2008; Newman, Pennebaker, Berry, and Richards, 2003), in courtroom transcripts of those found guilty of committing a crime and convicted of perjury (Huddle & Pennebaker, 2009), or in online dating profiles (Toma & Hancock, 2012), language analyses have shown that there is less use of first person singular pronouns in deceptive statements. Presumably, this is due to the lack of ownership of deceptive statements or psychological distancing.

However, in verified fake hotel reviews, it has been found that first person singular pronouns appear at higher rates, relative to genuine hotel reviews (Ott, Cardie, & Hancock, 2012). The authors theorize that placing one's self in the hotel setting is an important feature of this particular type of deception, suggesting that context is important when considering certain types of lies.

Word use has been tracked at the individual level to predict violent crimes by the Boston Bombers (Norman-Cummings & Pennebaker, 2013), by extremist groups in the Middle East (Pennebaker, 2011b), and by leaders intending on going to war (Chung & Pennebaker, 2011). In each of these cases, a drop in first person singular pronouns preceded violent acts. Since an attack involves hiding or concealing one's intentions to surprise "the enemy," it makes sense that language markers of deception are found in the language of attackers leading up to their violent acts.

From a forensics perspective, the ability to spot betrayal and secret-keeping has long been of interest to language scientists. For example, Niculae and colleagues (2015) found that it was possible to identify whether someone would betray their online gaming partner through more positive words, more politeness, and fewer words indicating future plans. Some cues for betrayal came from changes in language use by the victim; victims tended to increase in their use of planning words before betrayal. What was particularly intriguing was that the linguistic shifts in betrayal were apparent in both the betrayer and the betrayed through an increasing imbalance in the use of specific word categories between the interactants.

In another study, our research team tracked the emails of 62 people who admitted to keeping a major life secret from others (Tausczik, Chung, & Pennebaker, 2016). Participants were recruited online and were carefully screened in ways that preserved their anonymity, as well as the specific details of their secret. In all cases, those who agreed to release their previous year's emails were keeping secrets that, if discovered, would have been devastating to their lives or to the lives of others around them. As with the Niculae study, both the language of the secret-keepers and the language of the targets of the secrets changed. By examining function words, we were able to uncover the common ways in which people experience secret-keeping, and how it affected their relationships. Together, the results showed how psychological features can still be extracted when the topic of the exchange is not known to researchers. The secrets study in particular was the first to show how the language of a social network changes in response to a devastating life secret.

Even in forensic studies not involving extreme acts or violent crimes, language has provided clues to delinquent or mysterious activity, for example, in the papers of Diederik Stapel, a psychologist who was found to have been fabricating data for several of his published papers. An analysis of Stapel's articles confirmed to be fabricated vs. those not found to be fabricated showed more terms associated with scientific methods and certainty, and fewer adjectives (Markowitz & Hancock, 2014). These indicate that Stapel was emphasizing the novelty and significance of the results within the scientific literature, but unable to describe more concretely what he was reporting on.

Finally, investigators occasionally seek to learn who may have written a ransom note or even an entire book or play. As early as the 1960s, two statisticians applied early Bayesian analyses on *The Federalist Papers* to identify the authorship of a select group of disputed pamphlets. Mosteller and Wallace (1963) found that differences in a group of function words could serve as fingerprints to identify if the disputed papers were by Alexander Hamilton or James Madison. Similarly, Boyd and Pennebaker (2015) used LIWC and machine learning methods on both function and content words and concluded that a long-disputed play, *Double Falsehood*, was probably written by William Shakespeare.

The field of forensics will undoubtedly expand considerably in the years to come with increasingly sophisticated text analytic methods. Not only will investigators be able to identify authors but they will be able to better detect the intent or ongoing behaviors and personalities of the authors.

Health

When people are sick, uncertain about medical procedures, or dealing with health-related life changes, they inadvertently broadcast their situation online in ways potentially detectable by text analysis. For example, it is now possible to predict when couples might be expecting a child, and whether or not a new mother is likely to experience postpartum depression based on her tweets (de Choudhury, Counts, Horvitz, & Hoff, 2014). It is also possible to identify those at suicide risk from Facebook posts (Wood, Shiffman., Leary, & Coppersmith, 2016), or if individuals who are more or less likely to lose weight based on their diet blogs (Chung, 2009). Through other Twitter analyses, it is possible to isolate particular geographical locations more likely to experience higher rates of HIV (Ireland, Schwartz, Chen, Ungar, & Albarracín, 2015) or heart disease (Eichstaedt et al., 2015). Particularly exciting have been studies that track Wikipedia searches (Tausczik, Fasse, Pennebaker, & Petrie, 2012) or even vaping rates based on searches (Ayers et al., 2016). Note, however, that there have been failures to replicate some patterns gleaned from dynamic big data without corresponding traditional study methods (see Lazer, Kennedy, King, & Vespignani, 2014), such as flu epidemics based on Google searches (Ginsberg et al., 2009), suggesting that text analysis, like any other method in the social sciences, works best as a part of a multi-method toolkit.

Notice that these larger, more sociological or epidemiological questions appear more in content words, while clues to the more psychological questions appear more in function words. For example, in the aftermath of 9/11, livejournal.com blogs were examined for words representing preoccupation with the attacks (e.g., Osama, hijack, World Trade Center, etc.), positive and negative emotion words, and psychological distancing (i.e. a statistically derived index made of articles, first person singular pronouns, and discrepancy terms; Cohn, Mehl, &

Pennebaker, 2004). Not surprisingly, the analysis of content words revealed that communities across America were attending to death, religion, and the attacks much more after 9/11 than before. Mood went back to baseline levels (i.e. pre-9/11 levels) within a couple weeks, even for those who were highly preoccupied with the attacks. Function words, on the other hand, showed that psychological distancing persisted at least six weeks after 9/11. Individuals were talking less about themselves, and using "we" more to refer to their communities (Pennebaker, 2011). The content and function word analyses provided a timeline of topics and how widespread communities were psychologically responding to a massive upheaval in a naturalistic way.

The ability to detect and to predict the symptoms of various diseases, well-being, and community resilience after widespread upheaval is now possible through the text analysis of social media. Given that more and more of our interactions are online, and the ability to find people with similar symptoms and diseases has been made easier, it is possible that treatments, coping, recovery, and prevention strategies can be developed from our online interactions and behavior, although these are not without serious considerations, such as privacy, in their application (de Choudhury, 2013; Resnick, Resnick, & Mitchell, 2014; Wood et al., 2016).

Culture

Words can mark changes over time and place, providing new ways to assess the attentional focus of individuals, groups, and entire societies. By aggregating texts from the historical record, we can begin to track large-scale historical and cultural trends. The largest project of words over time examined keywords across 4 million digitized books (Michel et al., 2011). The authors counted word use over time to assess cultural trends (e.g., sushi, plagues, technology, etc.). The text analysis of cultural products has also allowed for the examination of psychological trends over time, including the examination of various values (Bardi, Calogero, & Mullen, 2008), individualism vs. collectivism (Twenge, Campbell, & Gentile, 2012), and sentiment (de Wall, Pond, Campbell, & Twenge, 2011) across recent history. Custom dictionaries using LIWC have been used to track possible cultural differences in the moral foundations of Liberals and Conservatives (Graham, Haidt, & Nosek, 2009), and in the relationships between prosocial language as predictive of public approval of U.S. Congress (Frimer, Aquino, Gebauer, Zhu, & Oakes, 2015).

Even within smaller geographic regions, it has been possible to track how pronoun use over time is associated with inciting action, for example, rallying for a revolution in the lead up to Iranian elections (Elson, Yeung, Roshan, Bohandy, & Nader, 2012), and within an online community, how pronouns are indicative of community tenure or life stage (i.e., number of posts from joining to leaving; Danescu-Niculescu-Mizil, West, Jurafsky, Leskovec, & Potts, 2013).

Given the access to digital written pieces, collaborative works, and interactions taking place over global platforms, there has been a growth in the assessment of regional differences in social psychological processes (Rentfrow, 2014). For example, a text analytic study of a corpus of essays on American's beliefs was assessed for values held across various states, and their relationship to state-level statistics published by national agencies (Chung, Rentfrow, & Pennebaker, 2014). Similarly, studies of tweets across major cities in America showed relationships of word use to rates of heart disease (Eichstaedt et al., 2015) and HIV prevalence (Ireland et al., 2015) assessed by the Centers for Disease Control.

Another study of tweets by U.S. counties found that high state level well-being and life satisfaction was associated with words indicating outdoors activities, spiritual meaning, exercise, and good jobs; low state life satisfaction was associated with negative emotion words indicating boredom (Schwartz et al,. 2013b). A study of Facebook posts suggested that relative positive and negative emotion word use could form an unobtrusive assessment of gross domestic happiness (Kramer, 2010). Together, these studies suggest that word comparisons across geographies can reveal systematic social processes that indicate relative health or well-being, providing insights into the relationships between sociological forces on psychological processes.

The Promise of Text Analytic Methods

Text Analytic Goals Across Fields

There are special tools, such as TACIT (Dehghani et al., 2016) that amass, extract, and code language. With word counts as the basic foundation of quantitative text analysis, a variety of simple and highly complex statistical techniques have been applied to code and decode text. For example, one software tool to derive psychological insights from word counts of content words is the Meaning Extraction Helper (Boyd, 2016), which facilitates the implementation of the Meaning Extraction Method (Chung & Pennebaker, 2008), a factor analysis of words to inductively extract themes from text. Note that there are a growing number of open-vocabulary approaches to analyze text (Schwartz et al., 2013a).

Text analysis methods span multiple disciplines. The interested reader should explore computer science and linguistics and, within both fields, natural language processing (NLP). NLP has the goal of classifying documents according to their features, and specifically, by the language used within a document. NLP methods are useful for psychological quantitative text analysis, but differ from psychological methods in various ways.

NLP methods typically require more programming skills to implement than are offered in traditional social psychology graduate programs, with much more preprocessing of text involved. There are a wide variety of open-source toolkits (e.g., Natural Language Toolkit [or NLTK], Stanford CoreNLP, etc.) that provide the code to execute the myriad of preprocessing steps (e.g., stemming, lemmatizing, tokenization, etc.) to prepare text for feature extraction, as well as to carry out a variety of analysis.

A very general distinction between social psychologists and NLP researchers is their purpose for studying language. Social psychologists typically use language as a reflection of social, cognitive, or emotional processes of the speaker or writer. NLP scientists, on the other hand, have one of two general underlying motives for their interest in language. Linguists or computational linguists typically analyze language because they tend to be more interested in language, and typically (but not always) are less interested in context or the attributes of the speaker than are psychologists. In contrast, computer scientists use language to categorize attributes into two or more categories using a variety of statistical methods. For example, it is possible to distinguish between genuine emails from spam by analyzing the features of the emails themselves. Like linguists, computer scientists are generally less interested as psychologists are in learning about the social psychological dynamics of the author of the texts themselves.

Technological Enablers of Text Analysis Growth

Particularly exciting is that researchers from psychology, computer science, and linguistics are now beginning to work together as part of a new discipline variously called Cognitive Science, Computational Social Science, and/or Artificial Intelligence. Each of these fields has benefited from the ways in which humans communicate, work, and socialize, resulting in significant scientific steps forward. The increase in CMC has enabled us to record, amass, extract, code and decode, and assess the meaning and style of text. These, in turn, have enabled us to develop more insights into potential applications with which to communicate digitally, or to capture communication digitally. Along with the increase in connectivity and CMC, there has been a surge of work in real-time speech-to-text capabilities, more language based digital art, machine translation, optical character recognition, machine translation, faster computing, multimedia systems, and an internet of things, statistical learning techniques, and cross disciplinary and cross industry collaborations to support CMC, and accordingly, to support the analysis of natural language. The amount of data collected on a person in association with their communication patterns presents many opportunities for research and innovation.

For example, natural language samples from social media can typically be associated with the user or other meta-data available (geolocation, topic, group affiliation, time of post, etc.). When the information is public, such as Reddit posts or Twitter posts, there are APIs for extracting the information, or more user-friendly tools to call on the platform's API (e.g., TACIT, Dehghani et al., 2016). What does this mean for social psychologists?

Social Psychological Applications of Text Analysis

Language is a defining feature of human culture and we are now only beginning to be equipped with the tools to study it on a massive scale. With all the technological advancements, cross-disciplinary collaborations, and our greater reliance on CMC throughout all areas of our lives, our insights will only continue to grow faster, more creative, and with broader applications.

There are two important caveats. The first is that while the study of the data we create as we go about our daily lives brings great benefits to our understanding of ourselves, our relationships, our health, our work, our deviance, and our culture, there must be limits and controls to ensure these benefits are weighed against the costs. There will be greater focus on the use of data and opt-outs/opt-ins beyond its collection, and terms of service or end-user license agreements (Kosinski, Matz, Gosling, Popov, & Stillwell, 2015; Mundie, 2014; PCAST, 2014; Verma, 2014).

The second caveat is that the variance that word scientists account for is low. We often publish exciting results because we have access to giant data sets. Enabled by advances in computerized text analysis and access to massive archives of digitized text, we are finding patterns that no one has seen before. But the effects are subtle. These insights into human behavior and social dynamics are illuminating, and are particularly handy when text is the only behavior we are able to objectively observe. However, their reliability and validity across contexts have yet to be assessed.

As our lives become increasingly digital, and the more we are able to extract psychological patterns from text, the more we open up to possibilities of being able to feedback analytics in real-time, or to predict behaviors across our social networks. We have never before been able to grasp what and whom we influence and how we influence to the degree that is possible today, and that possibility is only growing. The future of text analysis is amazingly exciting, with great potential for our understanding of how we are influenced by the actual, implied, or imagined presence of others.

References

Ayers, J. W., Althouse, B. M., Allem, J-P., Leas, E. C., Dredze, M., & Williams, R. S. (2016). Revisiting the rise of electronic nicotine delivery systems using search query surveillance. *American Journal of Preventive Medicine*, *40*(4), 448–453. doi:10.1016/j.amepre.2015.12.008

Bardi, A., Calogero, R. M., & Mullen, B. (2008). A new archival approach to the study of values and value-behavior relations: Validation of the value lexicon. *Journal of Applied Psychology*, *93*, 483–497.

Boyd, R. L. (2016). *Meaning extraction helper (MEH)*. Software program.

Boyd, R. L., & Pennebaker, J. W. (2015). Did Shakespeare write Double Falsehood? Identifying individuals by creating psychological signatures with text analysis. *Psychological Science*, *26*(5), 570–582. doi:10.1177/0956797614566658

Chung, C. K. (2009). *Predicting weight loss in diet blogs using computerized text analysis* (Unpublished dissertation). Austin, TX: University of Texas.

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

Chung, C. K., & Pennebaker, J. W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of Research in Personality*, *42*, 96–132.

Chung, C. K., & Pennebaker, J. W. (2011). Using computerized text analysis to assess threatening communications and actual behavior. In C. Chauvin (Ed.), *Threatening communication and behavior: Perspectives on the pursuit of public figures* (pp. 3–32). Washington, DC: The National Academies Press.

Chung, C. K., & Pennebaker, J. W. (forthcoming). What do you know when you LIWC a person? Text analysis as an assessment tool for traits, personal concerns, and life stories. In T. Shackelford & V. Ziegler-Hill (Eds.), *The* SAGE *handbook of personality and individual differences*. New York, NY: SAGE Publishing.

Chung, C. K., Rentfrow, P. J., & Pennebaker, J. W. (2014). Finding values in words: Using natural language to detect regional variations in personal concerns. In P. J. Rentfrow (Ed.), *Geographical psychology: Exploring the interaction of environment and behavior* (pp. 195–216). Washington, DC: The American Psychological Association.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science*, *15*, 687–693.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. L. (2012). *Echoes of power: Language effects and power differences in social interaction*. Proceedings of the 21st international conference on World Wide Web (WWW '12). New York, NY: ACM, pp. 699–708. DOI=http://dx.doi.org/10.1145/2187836.2187931

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013). *No country for old members: User lifecycle and linguistic changes in online communities*. Proceedings of the International World Wide Web Conference Committee (IW3C2), Rio de Janeiro, Brazil.

De Choudhury, M. (2013). *Role of social media in tackling challenges in mental health*. Proceedings of the 2nd international workshop on Socially-aware multimedia (SAM '13). New York, NY: ACM, 49–52. DOI=http://dx.doi.org/10.1145/2509916.2509921

De Choudhury, M., Counts, S., Horvitz, E., & Hoff, A. (2014). *Characterizing and Predicting Postpartum Depression from Facebook Data*. Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW '14). New York, NY: ACM, 626–638. DOI: https://doi.org/10.1145/2531602.2531675.

Deghani, M., Johnson, K. M., Garten, J., Boghrati, R., Hoover, J., Balasubramanian, V., Parmar, N. J. (2016). TACIT: An open-source text analysis, crawling, and interpretation tool. *Behavior Research Methods*, *49*(2), 538–547.

deWall, C. N., Pond, R. S., Jr., Campbell, W. K., & Twenge, J. M. (2011). Tuning in to psychological change: Linguistic markers of psychological traits and emotions over time in popular U.S. song lyrics. *Psychology of Aesthetics, Creativity, and the Arts*, *5*, 200–207.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., . . . Seligman, M. E. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, *26*(2), 159–169.

Elson, S. B., Yeung, D., Roshan, P., Bohandy, S. R., & Nader, A. (2012, February 29). *Using social media to gauge Iranian public opinion and mood after the 2009 election*. Santa Monica, CA: RAND Corporation, TR-1161-RC, 2012. Retrieved from www.rand.org/pubs/technical_reports/TR1161

Frimer, J. A., Aquino, K., Gebauer, J. E., Zhu, L., & Oakes, H. (2015). A decline in prosocial language helps explain public disapproval of the US Congress. *Proceedings of the National Academy of Sciences*, *112*, 6591–6594. doi:10.1073/pnas.1500355112

Galton, F. (1884). Measurement of character. *Fortnightly Review*, *36*, 179–185.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014.

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

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*, 1029–1046.

Hancock, J. T., Beaver, D. I., Chung, C. K., Frazee, J., Pennebaker, J. W., Graesser, A. C., & Cai, Z. (2010). Social language processing: A framework for analyzing the communication of terrorists and authoritarian regimes. *Behavioral Sciences in Terrorism and Political Aggression, Special Issue: Memory and Terrorism*, *2*, 108–132.

Hancock, J. T., Curry, L., Goorha, S., & Woodworth, M. T. (2008). On lying and being lied to: A linguistic analysis of deception. *Discourse Processes*, *45*, 1–23.

Hart, R., & Carrol, C. E. (2014). *Diction 7.0* [Computer software] Austin, TX: Digitext, Inc.

Huddle, D., & Pennebaker, J. W. (2009). *Language analysis of jury testimony from properly and wrongly convicted people*. Unpublished manuscript. University of Texas.

Ireland, M. E., & Mehl, M. R. (2014). Natural language use as a marker of personality. In T. Holtgraves (Ed.), *Oxford handbook of language and social psychology*. New York, NY: Oxford University Press.

Ireland, M. E., & Pennebaker, J. W. (2010). Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of Personality and Social Psychology*, *99*, 549–571.

Ireland, M. E., Schwartz, H. A., Chen, Q., Ungar, L., & Albarracín, D. (2015). Future-oriented Tweets predict lower county-level HIV prevalence in the United States. *Health Psychology*, *34*, 1252–1260.

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*, *22*(1), 39–44. doi:10.1177/0956797610392928

Kacewicz, E. (2013). *Language as a marker of CEO transition and company performance* (Unpublished dissertation). Austin, TX: The University of Texas at Austin.

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2013). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, *33*, 124–143. doi:10.1177/0261927X1350265

Kosinski, M., Matz, S., Gosling, S., Popov, V., & Stillwell, D. (2015). Facebook as a social science research tool: Opportunities, challenges, ethical considerations and practical guidelines. *American Psychologist*, *70*(6), 543.

Kramer, A. D. I. (2010). *An unobtrusive behavioral model of "gross national happiness"*. Proceedings of Computer-Human Interaction (CHI), pp. 287–290.

Kramer, A. D. I., & Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 8788–8790. doi:10.1073/pnas.1320040111

Laserna, C. M., Seih, Y., & Pennebaker, J. W. (2014). Um. who like says you know: Filler word use as a function of age, gender, and personality. *Journal of Language and Social Psychology*, *33*, 328–338. doi:10.1177/0261927X14526993

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014, March 14). The parable of Google Flu: Traps in big data analysis. *Science*, *343*(6176), 1203–1205. doi:10.1126/science.1248506

Markowitz, D. M., & Hancock, J. T. (2014). Linguistic traces of a scientific fraud: The case of Diederik Stapel. *PLoS ONE*. Retrieved from http://dx.doi.org/10.1371.journal.pone.0105937

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

Mergenthaler, E. (1996). Emotion-abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes. *Journal of Consulting and Clinical Psychology*, *64*, 1306–1315.

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

Michel, J-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, . . . Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *14*, 176–182. doi:10.1126/science.1199644

Mosteller, F., & Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *Journal of the American Statistical Association*, *58*, 275–309.

Mundie, C. (2014). Privacy pragmaticism: Focus on data use, not data collection. *Foreign Affairs*, *93*(2), 28–38.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic style. *Personality and Social Psychology Bulletin*, *29*, 665–675.

Niculae, V., Kumar, S., Boyd-Graber, J., & Danescu-Niculescu-Mizil, C. (2015). *Linguistic harbingers of betrayal: A case study on an online strategy game*. Proceedings of the Association for Computational Linguistics (ACL2015). 1. 10.3115/v1/P15-1159.

Norman-Cummings, B., & Pennebaker, J. W. (2013). *Tracking the Tweets of the Boston Marathon Bomber: A text analysis strategy for threat detection*. Unpublished manuscript, University of Texas.

O'Connor, B., Bamman, D., & Smith, N. A. (2011). Computational text analysis for social science: Model assumptions and complexity. *Public Health*, *41*: 43.

Ott, M., Cardie, C., & Hancock, J. (2012). *Estimating the prevalence of deception in online review communities*. Proceedings of the International World Wide Web Conference Committee (pp. 201–210). ACM.

Pennebaker, J. W. (2011a). *The secret life of pronouns: What our words say about us*. New York, NY: Bloomsbury Press.

Pennebaker, J. W. (2011b). Using computer analyses to identify language style and aggressive intent: The secret life of function words. *Dynamics of Asymmetric Conflict*, *4*, 92–102. Retrieved from http://dx.doi.org/10.1080/17467586.2011.627932

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic inquiry and word count (LIWC2015)*. Austin, TX. Retrieved from www.liwc.net

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic inquiry and word count (LIWC2007) [Computer software.]*. Austin, TX. Retrieved from www.liwc.net

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin. doi:10.15781/T29G6Z

Pennebaker, J. W., & Chung, C. K. (2008). Computerized text analysis of al-Qaeda statements. In K. Krippendorff & M. Bock (Eds.), *A content analysis reader* (pp. 453–466). Thousand Oaks, CA: Sage Publications.

Pennebaker, J. W., Chung, C. K., Ireland, M. I., Gonzales, A. L., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007*. Austin, TX. Retrieved from liwc.net

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count (LIWC2001)* [Computer software]. Mahwah, NJ: Lawerence Erlbaum Associates.

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

President's Council of Advisors on Science and Technology (PCAST; 2014). *Report to the President: Big data and privacy: A technological perspective*. Retrieved from www.whitehouse.gov/ostp/pcast

Ranganath, R., Jurafsky, D., & McFarland, D. (2009). *It's not you, it's me: Detecting flirting and its misperception in speed-dates*. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Association for Computational Linguistics, pp. 334–342.

Rentfrow, P. J. (2014). *Geographical psychology: Exploring the interaction of environment and behavior*. Washington, DC: The American Psychological Association.

Resnick, P., Resnick, R., & Mitchell, M. (2014). *Workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*. Proceedings of the 2014 Association for Computational Linguistics. Stroudsberg, PA: Association for Computational Linguistics.

Robbins, M. L., Mehl, M. R., Holleran, S. E., & Kasle, S. (2011). Naturalistically observed sighing and depression in rheumatoid arthritis patients: A preliminary study. *Health Psychology*, *30*, 129–133.

Romero, D. M., Swaab, R. I, Uzzi, B., & Galinsky, A. D. (2015). Mimicry is presidential: Linguistic style matching in presidential debates and improved polling numbers. *Personality and Social Psychology Bulletin*, *41*(10), 1311–1319.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Agrawal, M., Park, G. J., & Lucas, R. E. (2013b). *Characterizing geographic variation in well-being using tweets*. In Seventh International AAAI Conference on Weblogs and Social Media (ICWSM 2013), online. Retrieved from http://wwbp.org/papers/icwsm2013_cnty-wb.pdf

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, Shah, A., . . . Ungar, L. H. (2013a). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*. Retrieved from http://dx.doi.org/10.1371/journal.pone.0073791

Slatcher, R. B., & Pennebaker, J. W. (2006). How do I love thee? Let me count the words: The social effects of expressive writing. *Psychological Science*, *17*, 660–664.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis*. Cambridge, MA: MIT Press.

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016). *Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions*. Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, pp. 613–624.

Tausczik, Y. R., & Chung, C. K., & Pennebaker, J. W. (2016). *Tracking secret-keeping in emails*. Proceedings of the 2016 International Conference on Weblogs and Social Media, (pp. 388–397).

Tausczik, Y. R., Fasse, K., Pennebaker, J. W., & Petrie, K. J. (2012). Public anxiety and information seeking following the H1N1 outbreak: Blogs, newspaper articles, and Wikipedia visits. *Health Communication*, *27*, 179–185.

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

Toma, C. L., & Hancock, J. T. (2012). What lies beneath: The linguistic traces of deception in online dating profiles. *Journal of Communication*, *62*, 78–97.

Twenge, J. M., Campbell, W. K., & Gentile, B. (2012). Increases in individualistic words and phrases in American books, 1960–2008. *PLoS ONE*, 7(7), e40181.

Verma, I. M. (2014). PNAS Editorial expression of concern and correction. *PNAS*, *111*(29), 10779.

Wood, A., Shiffman, J., Leary, R., & Coppersmith, G. (2016). *Language signals preceding suicide attempts*. Proceedings of Computer-Human Interaction (CHI 2016), San Jose, CA.