Counting Little Words in Big Data:

The Psychology of Communities, Culture, and History

Cindy K. Chung and James W. Pennebaker

The University of Texas at Austin

Correspondence should be addressed to Cindy K. Chung (CindyK.Chung@mail.utexas.edu) or James W. Pennebaker (Pennebaker@mail.utexas.edu).

Counting Little Words in Big Data:

The Psychology of Communities, Culture, and History

Language can provide a window into individuals, families, and their community and culture, and, at the broadest level into history. Words are the primary means by which we express our thoughts and feelings. They are what we use to communicate and archive our experience of events. Given the centrality of language, it is somewhat surprising that so few social scientists have relied on word analyses to understand basic social processes. The reason, of course, is that until the very end of the 20th century, large-scale word analyses were simply too difficult to do. With the simultaneous popularity of the desktop computer and the internet, researchers for the first time were able to explore natural language on a scale never imagined.

Our approach to language has been to count words in a number of grammatical, psychological, and content categories using a computerized software program. The program was initially developed to understand psychological processes in individuals who had provided language samples in lab and clinical studies. In recent years, it has become a widely used tool for linguistic and literary studies, and for the analysis of social media data sets on the scale of billions of words in many languages and across hundreds of centuries.

In this chapter, we begin by describing the development and initial applications of computerized text analysis programs in lab and clinical psychology studies. One pattern that continually arose in the first decade of these studies was that many psychological effects were associated with relative rates of function word use (Pennebaker, 2011). That is, much of the variance in language to identify psychopathologies, honesty, status, gender, or age, was heavily dependent on the use of little words such as articles, prepositions, pronouns, etc., more than on content words (e.g., nouns, regular verbs, some adjectives and adverbs). These patterns have

been replicated across a variety of data sets (see Chung & Pennebaker, 2012; Pennebaker, Mehl, & Niederhoffer, 2003; Tausczik & Pennebaker, 2010).

With growing archives of computer-mediated communication on the internet, along with curated archives within the information sciences and digital humanities, the potential uses for computerized text analysis methods have expanded beyond understanding the psychology of an individual. We review studies that have counted little words in big data to understand the psychology of communities, cultures, and history. Our review focuses on studies that have used a variety of other natural language processing methods to address social science research questions more than on heavily computational or linguistics research questions. We conclude with a discussion of how analyses of both lab and real-world archives of natural language together can inform our understanding of our selves, our cultures, and our history.

Background: The Development of A Computerized Text Analysis Program

Previous research has found that participants who write for 15 to 20 minutes about their deepest thoughts and feelings surrounding a negative or traumatic event for 3 to 4 consecutive days experience later improvements in mental and physical health relative to participants who write about non-emotional topics (Pennebaker & Beall, 1986). The effects of this experimental paradigm, termed expressive writing, have been replicated across dozens of studies, across labs, and in different countries (for reviews, see Frattaroli, 2006; Pennebaker & Chung, 2011). In an attempt to identify the salutary mechanisms of expressive writing, we developed a computer program to automatically count words relevant to psychological processes in the growing archive of hundreds of expressive writing essays.

To measure the degree to which participants were expressing emotions, lists were derived of words and synonyms denoting, say, positive emotions, such as *excited, happy, love, optimism,*

and *win,* and judges voted on whether or not each word belonged in that category. The same process was repeated for other psychologically relevant categories including cognitive mechanisms, social processes, and biological processes. Content categories such as home, death, religion, etc. were included to measure the degree to which various topics were being discussed. Finally, closed class words, otherwise known as function words, or junk words, were included since these are previously established categories in the English language and so they could easily be added to the dictionary.

Ultimately, the computer program, made up of a text processor and the aforementioned dictionary, was called Linguistic Inquiry and Word Count (LIWC2001; Pennebaker, Francis, & Booth, 2001). LIWC, pronounced "Luke," computes the percentages of words used by a speaker or author that are devoted to grammatical (e.g., articles, pronouns, verbs) and psychological (e.g., emotions, cognitive mechanism words, social words) categories. The entries and categories for the LIWC dictionary were revised in 2007 (LIWC2007; Pennebaker et al., 2007), with certain categories culled, created, or expanded. The processor, which matches words in the text that it processes to the dictionary that it references, remained largely the same in the 2007 revision, but with the ability to process Unicode text and phrases, and to highlight dictionary matches within the text. For a demo of LIWC, visit www.liwc.net.

LIWC was first applied to the expressive writing corpus to determine the degree to which word use along certain categories might be predictive of later improvements in health. Increases in cognitive mechanism words (e.g., *because, insight, realize, understand,* etc.) and positive emotion words, along with a moderate use of negative emotion words were found to predict later health (Pennebaker, Mayne, & Francis, 1997). More importantly, the LIWC analysis suggested that participants who were able to make realizations and find benefits from their experience,

while acknowledging the negative aspects of a negative event were more likely to experience improved health in the weeks after expressive writing. That is, LIWC was able to identify language markers for a variety of processes known to be associated with adaptive psychological coping. Counting words was an effective means by which to understand how an author was relating to their topic, with theoretically meaningful relationships to practically important outcomes.

I. Language as a Window Into the Individual Soul

The initial tests of the LIWC methodology suggested that the ways people used language could serve as a window into basic social and psychological processes. As outlined below, our lab and others soon discovered that the analysis of function words yielded a number of promising and oftentimes surprising effects.

Mood Disorders

LIWC was then applied as a tool to understand psychological processes in a variety of texts from lab and clinical studies, with some studies seeking convergent validity from online natural language samples. A recurrent finding was that the largest associations between language and other psychological measures were found in relative function word use. That is, rates of function word use showed stronger relationships to depression, bipolar disorder, and suicide than did other LIWC categories, including categories of emotion word use. For example, a relative increased rate of first person singular pronouns (e.g., *I, me, my*) has been found in the college essays of depressed students relative to non-depressed students (Rude, Gortner, & Pennebaker, 2004), in online bulletin board messages devoted to the discussion of depression relative to discussion of home improvement or dieting (Ramirez-Esparza, Chung, Kacewicz, & Pennebaker, 2008), and forum comments by those with bipolar disorder relative to loved ones searching for

information on the disorder (Kramer, Fussell, & Setlock, 2004). Across each of these studies, depression and bipolar disorder was characterized by self-focus more than on attention to negative topics and emotions.

Similar effects have been found for suicide. Suicide has been characterized by social isolation (Durkheim, 1951) and heightened self-focus (Baumeister, 1990). Indeed, in the analyses of the collected works by suicidal poets relative to non-suicidal poets (Stirman & Pennebaker, 2001), and in case studies of suicide completers (for a review, see Baddeley, Daniel, & Pennebaker, 2011), suicidal individuals tended to show increasing social isolation and heightened self-focus in their increasing rates of "I" use and decreasing rates of "we" use over time. Negative emotion use tends to increase approaching suicide, but with changes in positive emotion word use mostly limited to studies that examine short time frames (less than 1 to 2 years). Again, the links between suicide and pronoun use generally tend to be larger than the effects for emotion word use.

Personality and Demographics

Function words have also been found to be associated with various personality traits in archival experimental studies (see Mehl, Gosling, & Pennebaker, 2006), blogs (Oberlander & Nowson, 2006; Nowson & Oberlander, 2007; Yarkoni, 2010), text messages (Holtgraves, 2010), and instant messaging chats in virtual worlds (Yee, Harris, Jabon, & Bailenson, 2011). For a demo of personality in the twittersphere, visit www.analyzewords.com. Accordingly, function words play a large role in investigations of author attribution, such as age, sex, and social class (e.g., Argamon, Koppel, Pennebaker, & Schler, 2009). It has been found, for example, that women tend to use more personal pronouns relative to men, representing their greater attention to social dimensions. On the other hand, men tend to use more articles (i.e. *a, an, the*),

representing their greater attention to more concrete details (Newman, Groom, Handelman, & Pennebaker, 2008).

There is some evidence that the relative rates of pronoun use between men and women are associated with levels of the hormone testosterone. For example, in case studies of patients who were administered testosterone at regular intervals, rates of pronouns referring to others decreased in journal entries and emails immediately following the testosterone injections. Pronouns referring to others increased as testosterone levels dropped in the following weeks. These results suggest that testosterone may have the effect of steering attention away from others as social beings (Pennebaker, Groom, Loew, & Dabbs, 2003). Across each of these studies, it is important to note that while women and men may have varied in the kinds of topics they discussed, examining the relative rates of function word use is reliably informative of sex across a variety of topics.

II. Language as a Window Into Relationships

Relationship Quality

Function words can convey attention to others as social beings. Several studies have examined the degree to which function words are a marker of relationship quality and stability. Previous studies have shown that using we at high rates in interactive tasks in the lab predict relationship functioning and marital satisfaction (Gottman and Levenson, 2000; Sillars, Shellen, McIntosh, & Pomegranate, 1997; Simmons, Gordon, and Chambless, 2005). Another study found that we-use, reflecting a communal orientation to coping by spouses in interviews about a patient's heart failure condition was predictive of improvements in heart failure symptoms of the patient in the months following the interview (Rohrbaugh, Mehl, Shoham, Reilly, & Ewy, 2008). However, a study of over 80 couples interacting outside of the lab with each other via IM over

10 days failed to show a relationship with we. Rather, the more participants used emotion words in talking with each other – both positive and negative emotion words – the more likely their relationship was to survive over a 3 to 6 month interval (Slatcher & Pennebaker, 2006). The research suggests that although brief speech samples can be reliably related to the functioning and quality of a relationship, natural language outside of the lab can provide a different picture of what types of communication patterns are associated with long-term relationship stability

The Development of Intimate Relationships: Speed-Dating

Rather than looking at overall levels of function words, several studies have assessed the degree to which interactants use function words at similar rates, termed language style matching (LSM), is associated with relationship outcomes. For example, an analysis of speed-dating sessions showed that LSM could predict which of the interactions would lead to both parties being interested in going out on a real date (Ireland, Slatcher, Eastwick, Scissors, Finkel, & Pennebaker, 2011). The transcripts came from a series of heterosexual speed-dating sessions offered on the Northwestern University campus. Forty men and forty women participated in 12 four-minute interactions with members of the opposite sex. Following each interaction, participants rated how attractive and desirable the other person had been.

On the day following the speed-dating sessions, each person indicated whether or not they would be interested in dating each of the partners with which they had interacted. Both parties had to agree that they were interested in order for a "match" to occur, and only then were they given contact information to set up a potential date in the future. "Matches" were far more likely if LSM during the speed-dating interactions was above the median. Particularly interesting was that the LSM measures actually predicted successful matches better than the post-interaction

ratings of the individuals. In other words, LSM was able to predict if the couples would subsequently go out on a date better than the couples themselves.

In another corpus of three speed-dating sessions, Jurafsky, Ranganath, and McFarland (2009) analyzed 991 4-minute speed-dating sessions and found, among other dialogue and prosodic features, that judgments of speakers by daters as both friendly and flirting were correlated with the use of "you" by males, and "I" by females. They also found that men perceived by dates as awkward used significantly lower rates of "you". In another study on the same corpus, the authors (Ranganath, Jurafsky, & McFarland, 2009) found that the pronoun cues were generally accurate: men who reported flirting used more "you", and more "we" among other features; women who reported flirting used more "I" and less "we". Note that language analyses to detect self-reported intent to flirt were much better than daters' perceptions of their speed-date's flirting.

Instant Messages (IMs) and Other Love (and Hate) Letters

Whereas the speed dating project focused on strangers seeking partners, another project assessed whether LSM could also predict the long term success of people who were already dating. In a reanalysis of an older study (Slatcher & Pennebaker, 2006), the instant messages (IMs) between 86 heterosexual romantic couples were downloaded before, during, and after participation in a psychology study. LSM between the couples was computed over 10 days of IMs. Almost 80% of couples with high LSM (above the median) were still together three months later, whereas only half of the couples with low LSM (below the median) were together three months later. LSM was able to predict the likelihood of a romantic couple being together three months later over and above self-reported ratings of relationship stability.

LSM has also been applied to historical relationships based on archival records (Ireland & Pennebaker, 2010). The correspondence between Sigmund Freud and Carl Jung is famous in tracking their close initial bonds and subsequent feud and falling out. The sometimes passionate and sometimes tumultuous romantic relationships of Elizabeth Barrett-Browning and Robert Browning as well as Sylvia Plath and Ted Hughes were refered to in their poetry for years before the couples met, during the happy times of their marriage, and the less-than-happy. Across all cases, LSM reliably changed in response to times of relationship harmony (higher LSM) and in times of relationship disharmony (lower LSM). Interestingly, even without the use of self-reports, LSM was able to reliably indicate relationship dynamics over time. Since these language samples had been recorded for purposes other than assessing group dynamics, they provide evidence regarding the robustness of LSM to predict real world outcomes beyond a controlled laboratory study.

III. Language as a Window Into a Community

Talking On the Same Page: Wikipedia and Craigslist

The current generation of text analytic tools is allowing us to track ongoing interactions for the first time. Two venues that have been of particular interest have been Wikipedia and CraigsList. In both cases, hundreds of thousands of people contribute to these online sites leaving traces of their communication and social network patterns.

Wikipedia, which started in 2001, is an online encyclopedia-like information source that has more than 3 million articles. Many of the articles are written by experts on a particular topic and have been carefully edited by dozens, sometimes hundreds of people. For the most commonly-read articles, an elaborate informal review takes place. Often, a single person will begin an article on a particular topic. If it is a topic of interest, others will visit the site and

frequently make changes to the original article. Each Wikipedia article is a repository of group collaboration. The casual visitor only sees the current final product. However, by clicking on the "discussion" tab, it is possible to see archives of conversations among the various contributors.

Wikipedia discussions are a naturalistic record of interactions among the various editors of each article. Recently, the discussion threads of about 70 Wikipedia articles (all about American mid-sized cities) that had been edited multiple times by at least 50 editors over several years were analyzed (Pennebaker, 2011). By comparing the language of each entry, it is possible to calculate an overall LSM score. Wikipedia sponsors an elaborate rating system that categorizes articles as being exemplary, very good, good, adequate, or poor.

Across the 70 Wikipedia entries, the higher the LSM of the discussions, the higher the rating for the entry, $r\ (68) = .29$, $p < .05$. The LSM levels for discussion groups were quite low relative to other data sets, averaging .30 -- likely due to the highly asynchronous communication in Wikipedia discussions. Nevertheless, the highest, mid-level, and lowest rated articles had LSM coefficients of .34, .30, and .27, respectively. In other words, Wikipedia discussions that indicated that the editors were corresponding in more similar ways to each other tended to develop better products.

Whereas Wikipedia discussions involve minimally-organized communities of people interested in a common topic, it is interesting to speculate how broader communities tend to coalesce in their use of language. Is it possible, for example, to evaluate the overall cohesiveness of entire corporations, communities, or even societies by assessing the degree to which they use language within their broader groups?

As a speculative project, we analyzed CraigsList.com ads in 30 mid-size cities to determine if markers of community cohesiveness might correlate with language synchrony

(Pennebaker, 2011). During a month-long period in 2008, approximately 25,000 ads in the categories of cars, furniture, and roommates were downloaded. For each ad category, we calculated a proxy for LSM, the standard deviation of each of LSM's nine function word categories was computed by city and then averaged to build an LSM-like variability score (the psychometrics are impressive in that the more variability for one function word category, the greater the variability for the others – Cronbach alpha averages .75).

Overall, linguistic cohesiveness was related to the cities' income distribution as measured by the gini coefficient, $r(28) = .35$, $p = .05$. The gini statistic taps the degree to which wealth in a community is completely evenly distributed (where gini = 0) versus amassed in the hands of a single person (gini = 1.0). As can be seen in the table below, linguistic cohesiveness was unrelated to racial or ethnic distribution and to region of the country .

Table 2. Most and Least Linguistically Cohesive Cities in CraigsList Ads

Most Linguistically Cohesive Cities (Top 10)	**Least Linguistically Cohesive Cities (Bottom 10)**
Portland, Oregon	Bakersfield, California
Salt Lake City, Utah	Greensboro, North Carolina
Raleigh, North Carolina	Louisville, Kentucky
Birmingham, Alabama	Oklahoma City, Oklahoma
Rochester, New York	Dayton, Ohio
Hartford, Connecticut	El Paso, Texas
New Orleans, Louisiana	Jacksonville, Florida
Richmond, Virginia	Columbia, South Carolina
Worcester, Massachusetts	Tulsa, Oklahoma
Tucson, Arizona	Albany, New York

Note: Cohesiveness is calculated by the degree to which people in the various communities used function words at comparable levels.

The city-wide data is meant to be a demonstration of a possible application of a simple text analysis approach to understanding any group. In our view, LSM is reflecting the basic social processes in groups and communities. In other words, the analysis of function words may serving as a remote sensor of a group's internal dynamics.

Remotely Sensing Mood, Influence, and Status

While the previous studies examined group engagement, many studies have aimed to examine overall mood and influence within a community. For sentiment analysis, LIWC's positive and negative emotion word categories have been used to assess the relative positivity or negativity within an online forum (see Gill, French, Gergle, & Oberlander, 2008 for validation of the LIWC emotion word categories for sentiment analysis, particularly anger and joy, in blogs). For example, Chee, Berlin, and Schatz (2009) examined the use of LIWC's emotion word categories in Yahoo! Groups illness groups. They found expected changes in sentiment in response to FDA approval, media attention, withdrawal from the market, and remarketing of particular meds, suggesting that sentiment analysis could be used to examine how a market group feels and responds to a given product.

In social media sites, there are many forums in which previously unacquainted strangers are not aware of the reputations, expertise, or clout of its members. The archives of language in social media sites, then, provide records of how influence and status are established. Nguyen and colleagues (2011) used LIWC to compare LiveJournal bloggers with many vs. few friends, followers, and group affiliations. Bloggers with fewer friends, followers, and group affiliations used nonfluencies (e.g., *er, hmm, um*) and swear words (e.g., *ass, fuck, shit*) at high rates. On the other hand, bloggers with many friends, followers, and group membership used big words (i.e., words six letters or more) and numbers (e.g., *first, two, million*) at high rates. These results

suggest that more formality and precision in language style may be a feature of larger groups, whereas an informal style may limit an individual's popularity and influence in a social network.

Language also provides cues to status hierarchies in online communities. For example, in the analysis of emails between faculty, graduate students, and undergraduate students, it was shown that high status interactants tended to use more "we" and lower status interactants tended to use more "I" in their emails, suggesting greater self-focus by lower status interactants (Chung & Pennebaker, 2007). Similar effects have been found in other social media contexts such as online bulletin board message forums (Dino, Reysen, & Branscombe, 2009), and in instant messages between employees of a research and development firm (Scholand, Tausczik, & Pennebaker, 2010). Indeed, these pronoun effects were previously found to be robust across lab studies (Kacewicz, Pennebaker, Davis, Jeon, & Graesser, 2012), and in archival memos and documents (Hancock et al., 2010).

Beyond counts of function words, Danescu-Niculescu-Mizil and colleagues (2012) examined 240,000 Wikipedia discussions and found that lower status editors changed their language more (i.e. showed higher LSM) to match their higher status counterparts. Similar effects were reported in the same paper in an analysis of over 50,000 conversational exchanges in oral arguments before the U.S. Supreme Court, in which lawyers matched their language more to the Chief Justice than to Associate Justices. In other words, the social hierarchy within a community can be mapped by the use of function words, and especially through pronouns.

IV. Language as a Window Into a Culture

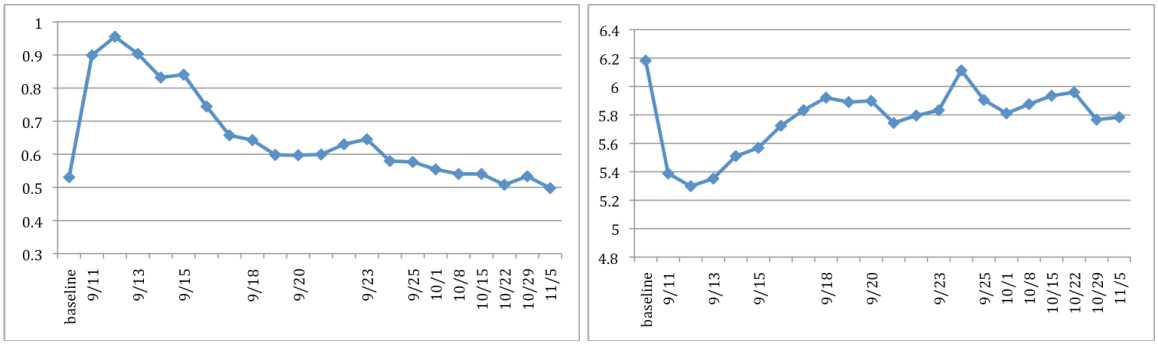
Shared Upheavals and Uprisings

The analysis of we-words (e.g., we, us, our) suggests that feelings of group identity are far more complicated than one might imagine. When appropriately primed, people naturally fuse

their identity with groups of importance to them. In classic experiments, Cialdini and his colleagues (1976) demonstrated that people were more likely to embrace their college football team's identity after a win than after a loss. This "we won" – "they lost" phenomenon was particularly strong if interviewed by people from another state than by people from their own community. Similarly, when groups are threatened from the outside, the usage of we-words increases dramatically.

Analyses of pronouns in 75,000 blog entries from about 1,000 bloggers in the weeks surrounding 9/11 demonstrated a dramatic and statistically significant jump in we-words and drop in I-words immediately after the terrorist attacks. These pronoun effects persisted in moderated form for up to a month after the attacks (reanalysis of Cohn, Mehl, & Pennebaker, 2001 data; in Pennebaker, 2011).

Figure 1. Pronoun Use by Bloggers Before and After September 11, 2001



Note. Graphs reflect percentage of we-words (left) and I-words (right) within daily blog entries of 1,084 bloggers in the two months surrounding September 11, 2001.

The use of social media has become an increasingly common real time news source in tapping how a culture responds to and anticipates events. Anecdotally, more and more people are turning to their Facebook wall and Twitter feeds for news on late-breaking events than to traditional news media such as newspapers and television. Social media as a news source for tracking events in different countries has been especially prevalent in the Arab spring, in terrorist

attacks, and in natural disasters, for which the experiences of citizens who may be inaccessible through traditional means, report on events in a local area. By analyzing the communications produced within a geographic location during a major event, it is possible to track the unfolding of thoughts, emotions, and behaviors of residents by the people who are experiencing it.

For example, Elson and colleagues (2012) analyzed over 2 million Iranian tweets in a 9 month period during the contested 2009 presidential elections until the end of protests in February 2010. “Twitter users sent tweets -- short text messages posted using Twitter -- marked with the “IranElection” hash tag (i.e., labeled as being about the Iran election) at a rate of about 30 new tweets per minute in the days immediately following the election.” The authors found that rates of LIWC’s swear words rose in the weeks leading up to protests. In addition, the rates of personal pronoun use, “I” and “you” in particular, were used at high rates in the protests immediately following the election and in leading up to one of the largest protests on September 18 (Quds Day in Iran). The use of these personal pronouns, a sign that people were focused on reaching out to others, evidenced a downward trend as the government instituted unprecedented crackdowns on protests beginning in October 2009. These findings show that little words can provide a window into how a culture is perceiving events and potentially, how they intend to respond.

Information and Misinformation

In addition to being a source of social connections, much of internet traffic is devoted to people searching for information. By analyzing where people go for information, we get a sense of their interests and concerns. Only recently have we begun to make the connection between emotional experiences and people’s need for specific types of information.

In late April, 2009, the World Health Organization announced the potential danger of a new form of flu, based on the H1N1 virus, more commonly known as the swine flu. Over the next 10 days, a tremendous amount of media attention and international anxiety was aroused. Using a new search system, Tausczik and colleagues (2012) identified almost 10,000 blogs that mentioned swine flu on a day by day basis. Analyses of the blogs revealed an initial spike in anxiety-related words that returned to baseline within a few days, followed by an increasing level of anger and hostility words. The authors further found that searching for information on Wikipedia tended to lag behind the swine flu mentions on blogs by about three days. These results suggest that after hearing about a potentially threatening disease, most of the public lets it stew for a few days before actively searching for information about its symptoms, time course, and treatment. Note that this strategy of information-seeking complements key word search strategies reported by Google and others (Ginsberg, Mohebbi, Patel, Brammer, Smolinski, & Brilliant, 2009) where online symptom searches actually lead diagnoses of flu across time and over regions.

Searching for information on the internet can also lead to misinformation. Accordingly, there is an increasing demand to identify misinformation on the internet, including SPAM (Drucker, Wu, & Vapnik, 2000), deceptive online dating profiles (Toma & Hancock, 2012), corporate scandal (Louwerse, Lin, Drescher, & Semin, 2010; Keila & Skillicorn, 2005), WikiCrimes such as Wikipedia vandalism (Harpalini, Hart, Singh, Johnson, & Choi, 2011), and deceptive product and service reviews (Ott, Choi, Cardie, & Hancock, 2011). By drawing on previous lab and forensic studies that had used LIWC to detect deception (see Hancock, Curry, Goorha, & Woodsworth, 2008; Newman, Pennebaker, Berry, & Richards, 2003), Ott and colleagues (2011) were able to develop algorithms to detect deceptive hotel reviews at rates well

above chance. For a demo, visit http://reviewskeptic.com/. Catching deviants and liars in an online community can be improved not just by the infrastructure of a given platform (e.g., SPAM guards, blocks, moderators, peer-rating systems, etc.), but by the ability to detect their linguistic fingerprints.

Sentiment Analysis: Is it Positive or Negative? And So What?

There has also been a growing interest within the field of natural language processing to characterize the sentiment of a culture. A growing number of computer scientists are interested in determining whether the overall mood within social media sites is relatively positive or negative, and then to predict various outcomes such as book sales (Gruhl, Guha, Kumar, Novak, & Tomkins, 2005), box office receipts (Mishne & Glance, 2006; Asur & Huberman, 2010), success in blogs devoted to weight loss (Chung, Jones, Liu, & Pennebaker, 2008), virality of news articles (Berger & Milkman, 2009), and stock market outcomes (Bollen, Mao, & Zeng, 2010; Gilbert & Karahalios, 2010). For a demo of mood in the twittersphere, visit http://www.ccs.neu.edu/home/amislove/twittermood/. For a demo of mood in the blogosphere, visit http://www.wefeelfine.org/.

Within the political realm, LIWC has been used to assess overall sentiment in congressional speeches as a step in classifying political party affiliation (Yu, Kaufmann, & Diermeier, 2008) . In addition, LIWC has been used to predict the outcome of Germany's 2009 federal elections from a sample of over one hundred thousand tweets (Tumasjan, Sprenger, Sandner, & Welpe, 2010).

Social psychologists have used LIWC to conduct sentiment analyses over time to characterize the prevalence of psychological constructs as a function of cultural events. deWall, Pond, Campbell, and Twenge (2011) found that rates of LIWC's positive emotion word use

decreased and rates of negative emotion word use increased from 1980 to 2007, which they claim are in line with other findings that rates of psychopathology, particularly narcissism and social disconnection, have increased over time. (There is some reason to question this parallel since narcissism is unrelated to pronoun use). In another study, Kramer (2010) used a dictionary-based system to assess gross national happiness across America in the status updates of 100 million Facebook users. By graphing a standardized metric of the difference in LIWC's positive and negative emotion word use across time, he found that Americans were more positive on national holidays (e.g., Christmas, Thanksgiving), and on the culturally most celebrated day of the week, Fridays. Kramer further found that Americans were the least positive on days of national tragedy (e.g., the day Michael Jackson died), and on Mondays. In other words, the dictionary-based metric was found to be a valid indicator of happiness as a function of the cultural context. For a demo of mood in Facebook, visit http://apps.facebook.com/usa_gnh/.

While the LIWC dictionary provides a previously validated measure of emotions, it should be emphasized that sentiment analysis provides only a small part of the big picture. Knowing the overall mood is informative of the degree to which a culture is celebrating, fearing, or angry about events. However, there are other little words that are just as easy to assess, and are much more telling of how an author, speaker, or group, is relating to their topic and to their social worlds. Pronouns tell us where and to whom people are paying attention (Pennebaker, 2011). Various prepositions tell us how complex or precisely people are thinking (Pennebaker & King, 1999). Auxiliary verbs tell us the degree to which expressions are story-like (Jurafsky et al., 2009). Going beyond sentiment analysis and analyzing function words allows us to remotely detect the social dynamics and thinking style of a culture.

V. Language as a Window Into History

Searching the Past for n-grams

Perhaps the largest scale analysis of cultural products has been the analysis of search terms (or n-grams, which are a continuous set of characters without spaces, in sets of *n*) in Google's digitized collection of 4% of all books ever published (Michel et al., 2011). The relative frequency of use of particular terms indicated the degree to which those terms were prevalent over the period 1800 to 2000, and therefore on the minds of individuals in the culture over time. For example, the authors examined the appearance of words indicating particular widespread diseases (e.g., *Spanish Flu*), cuisines (e.g., *sushi*), political regimes (e.g., *Nazis*), or religious terms (e.g., *God*) over time. Each of the terms rose and fell when the culture was experiencing change specific to the term. The authors termed this method of investigation "culturomics", which is a natural language processing method for highlighting cultural change (the concepts discussed), and linguistic change (the words used for a concept) in large corpora.

Following the culturomic approach, Campbell and Gentile (2012) examined trends in individualism and collectivism from 1960 to 2008. The authors examined the use of first person singular pronouns (e.g., *I, me, my*) and first person plural pronouns (e.g., *we, us, our*) using Google Ngram Viewer, which is an application that reports on the relative use of search terms in the Google Books Project over time. Presuming that "I" represents individualism and "we" represents collectivism, the authors found that there was a trend for increasing individualism and a decreasing trend for collectivism in English language books in the past half century. For a demo, try this yourself at http://books.google.com/ngrams. Note that this pattern of findings was also found in American popular song lyrics from 1980 to 2007 (de Wall et al., 2011).

Another approach to examine what has been on the culture's mind over time is to examine word categories that represent more topic relevant words. For example, Bardi and colleagues (2008) derived a lexicon of three words that typically tend to co-occur with each of Schwartz's Value Survey's ten categories of values. The lexicon was shown to be valid, with increases in their use in American newspapers during expected times across history (e.g., the words *power*, *strength*, and *control* to represent the Power value peaked in their collective occurrence in American newspapers during World War II, and was highly correlated with times of high military participation). Their study showed that lexicons of personal concerns can be used to examine the context in which those concerns are likely to be expressed, for example, during challenge or prosperity.

Conclusions

Social media sites are enabling the examination of social dynamics in unprecedentedly large samples. We are creating our own records of history simply by interacting as we naturally do -- by email, Facebook, Twitter, instant messaging (IM), text messages, etc. Accordingly, we have access to study our selves, our relationships, our communities, culture, and history through our own words. Since the turn of the century, a growing number of studies have used natural language processing methods to identify language patterns that signal even subtle psychological effects. Although some computing power, data mining, and database management are required for such large data sets, programs such as LIWC are easy to use, the dictionary that it references can be customized, and the results can easily be compared across studies. While lab and clinical studies are vital to understanding the psychology of individuals, counting little words in big data, just as has been found in smaller sample sizes, can shed light on the greater psychological context in which we communicate -- our communities, culture, and history.

On a broader level, the new language analysis methods have the potential to completely change the face of social psychology. By drawing on increasingly sophisticated computer-based methods on data sets from hundreds of millions of people, the traditional 2 x 2 laboratory methods of the 20th century begin to have an anachronistic feel. Indeed, the study of individuals and cultures can now be done faster, more efficiently, with far larger and more valid samples than has ever been possible.

In many ways, we view this work as a call to arms. If social psychologists want to exert a powerful influence on the acquisition of knowledge about groups and social dynamics, they must break from the past. By working with experts in social media, linguistics, communications, engineering, and the private sector, our discipline will become a central player in the social world. The failure to master these new technologies will result in our being co-opted by Google and other social media experts who desperately are trying to figure out social behavior in natural settings. Social psychologists of the world unite! We have nothing to lose but our complacency!

Acknowledgements

Department of Psychology A8000, University of Texas at Austin, Austin, Texas 78712. Correspondence should be addressed to CindyK.Chung@mail.utexas.edu or Pennebaker@mail.utexas.edu. Preparation of this manuscript was aided by funding from the Army Research Institute (W91WAW-07-C-0029) and National Science Foundation (NSCC-0904913). The authors would like to thank Mike Thelwall and Yla Tausczik for their helpful comments in the preparation of the manuscript.

Financial and Disclosure Issues: The LIWC2007 program, which is co-owned by Pennebaker, is commercially available for $89 USD (for the full package), $29 USD (student version), with discounts for bulk purchases on www.liwc.net. LIWC2007 demos, downloads, and products can be found on www.liwc.net. Text data for research purposes will be analyzed by Pennebaker free of charge. All profits that go to Pennebaker from LIWC2007 sales are donated to the University of Texas at Austin Psychology Department.

References

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the Association for Computing Machinery (CACM), 52*, 119-123.

Asur, S., & Huberman, B. A. (2010). *Predicting the future with social media*. arXiv:1003.5699v1

Baddeley, J. L., Daniel, G. R., & Pennebaker, J. W. (2011). How Henry Hellyer's use of language foretold his suicide. *Crisis, 32*, 288-292.

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.

Baumeister, R. F. (1990). Suicide as escape from self. Psychological Review, 97, 90?113.

Berger, J. A. and Milkman, K. L. (December 25, 2009). *What Makes Online Content Viral?* . Available at SSRN: http://ssrn.com/abstract=1528077 or http://dx.doi.org/10.2139/ssrn.1528077

Bollen, J., Mao, H., & Zeng, X.-J. (2010). Twitter mood predicts the stock market. Journal of *Computational Science, 2*, 1-8.

Campbell, W. K., & Gentile, W. (2012). *Cultural changes in pronoun usage and individualistic phrases: A culturomic analysis*. Talk presented at the 2012 Annual Meeting for the Society for Personality and Social Psychology, San Diego, CA.

Chee, B., Berlin, R., & Schatz, B. (2009). Measuring population health using personal health messages. In *Proceedings of the Annual American Medical Informatics Association (AMIA) Symposium*, 92-96.

Chung, C. K., Jones, C., Liu, A., & Pennebaker, J. W. (2008). Predicting success and failure in weight loss blogs through natural language use. *Proceedings of the 2008 International Conference on Weblogs and Social Media*, pp.180-181.

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. (2012). Linguistic Inquiry and Word Count (LIWC): pronounced "Luke" and other useful facts. In P. McCarthy & C. Boonthum, *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp.206-229). Hershey, PA: IGI Global.

Cialdini, R. B., Borden, R. J., Thorne, A., Walker, M. R., Freeman, S., & Sloan, L. R. (1976). Basking in reflected glory: Three (football) field studies. *Journal of Personality and Social Psychology, 34*, 366-375.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science, 15*, 687-93.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. In Proceedings of the 21st International World Wide Web Conference, 2012.

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.

Dino, A., Reysen, S., & Branscombe, N. R. (2009). Online interactions between group members who differ in status. *Journal of Language and Social Psychology, 28*, 85-93.

Drucker, H., Wu, D., & Vapnik, V. N. (2002). Support vector machines for spam categorization. *Neural Networks, IEEE Transactions on, 10(5)*: 1048-1054.

Durkheim, E. (1951). *Suicide*. New York: Free Press.

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

Frattaroli. J. (2006). Experimental disclosure and its moderators: A meta-analysis. *Psychological Bulletin, 132*, 823-865.

Gill, A. J., French, R. M., Gergle, D., & Oberlander, J. (2008). The language of emotion in short blog texts. In *Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work (CSCW)*, 299-302, San Diego, CA.

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*, 1012-1014.

Gottman, J. R., & Levenson, R. W. (2000). The timing of divorce: Predicting when a couple will divorce over a 14-year period. *Journal of Marriage and the Family, 62*, 737-745.

Gruhl, D., Guha, R., Kumar, R., Novak, J., & Tomkins, A. (2005). The predictive power of online chatter. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 78-87. New York: ACM Press.

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.

Harpalini, M., Hart, M., Singh, S., Johnson, R., & Choi, Y. (2011). Language of vandalism: Improving Wikipedia vandalism detection via stylometric analysis. *Association for Computational Linguistics (ACL2011)*.

Holtgraves, T. (2011). Text messaging, personality, and the social context. *Journal of Research in Personality, 45*, 92-99.

Ireland, M. E., & Pennebaker, J. W. (2010). Language style matching in reading and writing. *Journal of Personality and Social Psychology, 99*, 549-571.

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

Jurafsky, D., Ranganath, R., & McFarland, D. (2009). Extracting social meaning: identifying interactional style in spoken conversation. *Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT2009)*, 638-646.

Kacewicz, E., Pennebaker, J. W., Davis, D., Jeon, M., & Graesser, A. C. (2012). Pronoun use reflects standings in social heirarchies. Manuscript submitted for publication.

Keila, P.S. & Skillicorn, D.B. (2005). Detecting unusual email communication. *Proceedings of the 2005 Conference of the Centre for Advanced Studies on Collaborative Research (CASCON 2005)*, 238-246.

Kramer, A. D. I. (2010). An unobtrusive behavioral model of "gross national happiness". *Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 287-290. DOI=10.1145/1753326.1753369

Kramer, A. D. I., Fussell, S. R., & Setlock, L. D. (2004). Text analysis as a tool for analyzing conversation in online support groups. *Proceedings of the 23rd International Conference on Human Factors in Computing Systems (CHI2004)*: Late Breaking Results, pp. 1485-1488. New York: ACM Press.

Louwerse, M. Lin, K.-I., Drescher, A., & Semin, G. (2010). Linguistic cues predict fraudulent events in a corporate social network. *Proceedings of the 2010 Annual Meeting of the Cognitive Science Society*, 961-966.

Mehl, M.R., Gosling, S.D., & Pennebaker, J.W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology, 90*, 862-877.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 14*, 176-182. DOI:10.1126/science.1199644

Mishne, G., & Glance, N. (2006). Predicting movie sales from blogger sentiment. *Proceedings of the AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes, 45,* 211?246.

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.

Nguyen, T., Phung, D., Adams, B., & Venkatesh, S. (2011). Towards discovery of influence and personality traits through social link prediction. *Proceedings of the International Conference on Weblogs and Social Media (ICWSM2011)*. Barcelona, Spain.

Nowson, S., & Oberlander, J. (2007). Identifying more bloggers: towards large scale personality classification of personal weblogs. *Proceedings of the International Conference on Weblogs and Social Media (ICWSM2007)*.

Oberlander, J., & Nowson, S. (2006). Whose thumb is it anyway? Classifying author personality from weblog text. *Proceedings of the COLING Association for Computational Linguistics - Main Conference Poster Sessions*, 627-634.

Ott, M., Choi, Y., Cardie, C., & Hancock, J. (2011). Finding deceptive opinion spam by any stretch of the imagination. *Association for Computational Linguistics (ACL2011)*.

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

Pennebaker, J. W., & Beall, S. (1986). Confronting a traumatic event: Toward an understanding of inhibition and disease. *Journal of Abnormal Psychology, 95*, 274-281.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic Inquiry and Word Count (LIWC2007): A text analysis program*. Austin, TX: LIWC.net.

Pennebaker, J. W., & Chung, C. K. (2011). Expressive writing and its links to mental and physical health. In H. S. Friedman (Ed.), *Oxford handbook of health psychology* (pp.417-437). New York, NY: Oxford University Press.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count (LIWC; Version LIWC2001)* [Computer software]. Mahwah, NJ: Erlbaum.

Pennebaker, J.W., Groom, C.J., Loew, D., & Dabbs, J.M. (2004). Testosterone as a social inhibitor: Two case studies of the effect of testosterone treatment on language. *Journal of Abnormal Psychology, 113*, 172-175.

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

Pennebaker, J.W., Mayne, T.J., & Francis, M.E. (1997). Linguistic predictors of adaptive bereavement. *Journal of Personality and Social Psychology, 72*, 863-871.

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

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*, 334-342.

Ramirez-Esparza, N., Chung, C. K., Kacewicz, E., & Pennebaker, J. W. (2008). The psychology of word use in depression forums in English and in Spanish. Testing two text analytic approaches. *Proceedings of the 2008 International Conference on Weblogs and Social Media (ICWSM2008)*, pp. 102-108. Menlo Park, CA.

Rohrbaugh, M. J., Mehl, M. R., Shoham, V., Reilly, E. S., & Ewy, G. A. (2008). Prognostic significance of spouse "we" talk in couples coping with heart failure. *Journal of Consulting and Clinical Psychology, 76*, 781-789.

Rude, S. S., Gortner, E. M., & Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion, 18*, 1121-1133.

Scholand, A. J., Tausczik, Y. R., & Pennebaker, J. W. (2010). Social language network analysis. *Proceedings of Computer Supported Cooperative Work 2010*, pp. 23-26. New York: ACM Press.

Sillars, A., Shellen, W., McIntosh, A., & Pomegranate, M. (1997). Relational characteristics of language: Elaboration and differentiation in marital conversations. *Western Journal of Communication, 61*, 403-422.

Simmons, R. A., Gordon, P. C., & Chambless, D. L. (2005). Pronouns in marital interaction: What do you and I say about marital health? *Psychological Science, 16*, 932-936.

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.

Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine, 63*, 517-522.

Tausczik, Y., Faassee, K., Pennebaker, J. W., & Petrie, K. J. (2012). Public anxiety and information seeking following H1N1 outbreak: Blogs, newspaper articles, and Wikipedia visits. *Health Communication, 27*, 179-185. doi: 10.1080/10410236.2011.571759

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.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM2010)*, 178-185. Menlo Park, CA: AAAI Press.

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

Yee, N., Harris, H., Jabon, M., & Bailenson, J. N. (2011). The expression of personality in virtual worlds. *Social Psychological and Personality Science, 2*, 5-12.

Yu, B., Kaufmann, S., and Diermeier D. (2008). Classifying party affiliation from political speech. *Journal of Information Technology and Politics 5(1)*: 33-48.