## Title

Social Media and Political Movements: A Computational Exploration

Author
Chung, Justin

Publication Date
2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

# THE LANGUAGE OF PARTISANSHIP: A LINGUISTIC EXPLORATION OF POLITICAL BLOGS

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Informatics

by

Justin Chung

Thesis Committee:
Professor Gloria Mark, Chair
Professor Gary Olson
Professor Padhraic Smyth

2015

# DEDICATION

To my mom and dad.

## TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my committee chair, Professor Gloria Mark, who has spent countless hours to help me achieve this important milestone. I would like to thank my committee members Professor Gary Olson and Professor Padhraic Smyth for their helpful insights and patience throughout the process. I would also like to thank Graduate Dean Frances Leslie for her support. Thanks to Jen, Yiran, and Dakuo for their contributions and conversation throughout the early stages of this research and thanks to all those friends that attended my defense. I couldn’t have done it without you.

# CURRICULUM VITAE

## Justin Chung

justin.chung@uci.edu
40923 Arroyo Dr, Irvine, CA 92617

## EDUCATION

**University of California, Irvine**
2008-2015 Ph.D. Information and Computer Sciences, General track of the Informatics concentration; Advisor: Gloria Mark

**Stanford University**
2004–2008 B.S. Management Science & Engineering, graduated 2008. GPA 3.738/4.0; Advisor: Pamela Hinds

**Solon High School**, Solon, OH
2000-2004 Graduated 2004 - GPA 4.33/4.0, Valedictorian, Phi Beta Kappa, National Honor Society

## EXPERIENCE

**Graduate Student Researcher, Graduate Division, University of California, Irvine Irvine, CA**
2012-2013

Drove deployment of a CRM system for tracking and contacting prospective students, current students, and alumni. Gathered requirements, developed work processes and best practices, and conducted trainings.

**Research Intern, IBM Research — Collaborative User Experience Group Cambridge, MA**
Summer 2009

Worked on a study of enterprise community hosting applications. Conducted semi-structured interviews and used grounded theory to generate theory on the use of these applications.

**Research Assistant, Stanford Technology Ventures Program Stanford, CA**
Summer 2008

Worked with professor Tom Byers on a comprehensive revision of Byers, Dorf, and Nelson's *Technology Ventures: From Idea to Enterprise*. Added section on Web 2.0 enterprises and wrote over one hundred illustrative examples for book concepts.

**Research Assistant, Center for Work, Technology, and Organization**
**Stanford, CA**
2007-2008

Worked with advisor Pamela Hinds in a study on Human-Robot Interaction. Examined the role of various types of disclosure by robotic systems. Developed experimental procedure and manipulations. Coordinated and led team of 4 research assistants for lab setup and experimentation. Analyzed data using SPSS.

**Research Intern, Det Norske Veritas; Oslo, Norway**
Summer 2007

Researched the use and adoption of asynchronous groupware solutions in this multinational risk management consultancy as part of an ongoing research project. Used case study methods to research the relationship between aspects of work context and their impact on actual use of groupware. Gained experience in direct observation and interview as methods of data collection.

## PUBLICATIONS

Mark, G., Chung, J., Al-Ani, B., Jones, J.,"The Wikipedian Revolution: Collective Intelligence in the Egyptian Blogosphere." Collective Intelligence 2012.

Al Ani, B., Mark, G., Chung, J., Jones, J., "The Egyptian Blogosphere During the Revolution: A Narrative of Counter-Power." CSCW 2012. *Honorable Mention

Rosanne Siino, Justin Chung, Pamela Hinds. "Colleague vs. Tool: Effects of Disclosure in Human-Robot Interaction." IEEE RO-MAN 2008.

## SERVICE

**At-large Member, Board of Directors, UC Irvine Alumni Association**
2014-present

Helping to steer the strategic direction of the Alumni Association, helping develop plans for utilizing new technologies to deepen UCI alumni engagement with the Association.

**Vice President Financial Affairs, Associated Graduate Students, UC Irvine**
2014-2015

Served as Chief Financial Officer of graduate student government. Maintained travel grant and project funding programs, and created a new full-time professional staff position.

**Chair, Professional Development Subcommittee, Doctoral Student Support Steering Committee, University of California**
2014-2015

Member of committee composed of Academic Senate, administrative, and student government representatives, chartered by Provost Aimée Dorr, which sought to improve doctoral support in the interests of improving academic quality across the UC system. Served as one of four subcommittee chairs.

**Co-Chair, Council of Student Body Presidents, University of California**
2013-2014

Elected as leader of group consisting of the Student Body Presidents from all UC campuses. Strengthened ties with the University of California Office of the President and consulted on critical issues such as tuition, professional development, and mental health.

**Chair, Student Fees Advisory Committee, UC Irvine**
2012-2014

Advised the Provost / Executive Vice Chancellor on the use of UCI Student Services Fee revenue, which totaled over $24 million, intended for co-curricular programs and essential services. Revised the request process for permanent allocations to better track historical allocations and performance of funded programs. Reviewed and advised on adoption or rejection of course materials and services fees.

**President, Associated Graduate Students, UC Irvine**
2012-2014

Elected as President by student body. Managed and drove strategic planning and development of a student government organization whose purview includes campus programming, lobbying legislators, on-campus advocacy, and aspects of University/campus governance. Developed numerous programs, including a Symposium with eighty participants and $25,000 in prizes, a new travel grant program for student travel to conferences, and exponentially increased the number of events and social hours.

**Chair, Graduate and Professional Students Committee, University of California Students Association,**
2012-2014

Served as a student participant on the Graduate Division committee that produced the event welcoming all new PhD and Master's students to UCI. Gave a welcoming speech to all new graduate students.

**Vice President Financial Affairs, Associated Graduate Students, UC Irvine**

2011-2012

Served as Chief Financial Officer of the official graduate student government at UCI. Simplified chart of accounts to better meet organization's strategic needs, worked on issues like parking for graduate students, strengthened ties with the Pub, and maintained a balanced budget.

**Member,** New Graduate Student Orientation Steering Committee, UC Irvine, 2012-2014

**President,** Informatics Graduate Student Association, UC Irvine, 2009-2010

**Member,** Committee for Graduate Recruiting and Outreach, Department of Informatics, UC Irvine, 2009

**Student Volunteer**, ECSCW 2007, Dublin, Ireland.

**Student Volunteer**, CSCW 2006, Banff, AB

**Student Volunteer**, Stanford and CMU's People and Robots Workshop 2006, Carmel, CA

## HONORS AND AWARDS

**Graduate Research Fellowship,** National Science Foundation, 2010-2014

**Dean's Fellowship**, University of California, Irvine, 2008-2010

**Tau Beta Pi,** Stanford University. Honors society, top fifth of engineers by GPA, 2008

# ABSTRACT OF THE DISSERTATION

Social Media and Political Movements: A Computational Exploration

By

Justin Chung

Doctor of Philosophy in Informatics

University of California, Irvine, 2015

Professor Gloria Mark, Chair

In recent years, political polarization has received increased attention in the United States. Reports suggest that partisan differences have greatly deepened in the past several decades, and that this polarization has had a number of deleterious effects. However, some debate remains as to the etiology of political orientation and political choice—some suggest that political choices can be modeled as utility-optimizing economic actions, while others posit the existence of "moral foundations" that serve as the underpinnings of ideology. We take a quantitative approach to investigating these questions, examining over 80,000 posts from 50 liberal and 50 conservative blogs from the latter half of 2012, around the time of the United States presidential election. In our analysis, we employ Linguistic Inquiry and Word Count, a tool that measures the use of salient linguistic markers in digital texts. The markers examined include function words such as pronouns, articles, and prepositions; researchers have found that the frequency of usage of these words has correlations with a number of phenomena, including age, gender, wealth, and the success or failure of social interactions. Our investigation reveals that some divergence exists between the frequencies of use of these words by liberals and conservatives. By examining these divergences, we find support for a moral foundations approach to understand ideological differences. Our contribution adds to the discourse on the etiology of political choice and probes the ways in which ideological orientation affects written language.

# Chapter 1 : Introduction

## 1.1 Background

With the deepening ubiquity and cultural prominence of user-generated content delivered through social media, we are increasingly confronted with striking examples of how users of these media may express their views with unprecedented speed and to unprecedented audiences. Nielsen reports that in 2012, Internet users spent more time on social media sites than all other types of sites combined (Nielsen 2012). At the same time, Pew (Duggan and Smith 2013) reports that more than two-thirds of Americans are using some form of social networking. It is little wonder that the influence of social media in a variety of areas has been a topic of attention for many academics—these media have had indelible impacts on organizations, advertising, software engineering, education, and politics.

It is indisputable that the advent of social media has provided scientists of all stripes a unique opportunity to study large numbers of people producing original content in a quantity and variety never seen before. Those who study the patterns of behavior among people, organizations, and states have been presented with an organically generated, living record which could be analyzed in detail (Garton et. al. 1997). Scholars of communication have thoroughly explored the how and why of users' interaction with these "new media" (Kaplan and Haenlein 2010). In addition to communication, researchers across numerous disciplines have examined the relevance of social media to their fields—social media has had profound implications for medicine (Chou et. al. 2009), psychology (Bauer and Gaskell 2000), and education (Moran et. al. 2011) to name just a few examples.

As its influence and popularity has grown, social media has increasingly become a window through which researchers have sought to better understand American politics, both to understand the medium's role and influence in the political process and for insights into politics itself. Social media has inspired complete reimaginings of the public sphere (Shirky 2011) and citizen engagement (Dahlgren 2009), though some feel claims that democracy has fundamentally changed as a result of computer-mediated communication (CMC) are grandiose and unwarranted (Gladwell 2010).

Social media's use in the study of politics is not limited to making prognostications about its future, but these media also a valuable source of data for exploring its present. The sheer amount and variety of political discourse generated and published online is mind-boggling. It is with this stance that this dissertation proceeds—we seek to better understand the American political mind as it can be examined through the writings of politically-minded individuals.

## 1.2 Motivation

Over the past decade, an enormous amount of research in multiple disciplines have focused on understanding weblogs, or "blogs," and their authors. Blogs are defined from a technical perspective as “frequently modified web pages in which dated entries are listed in reverse chronological sequence,” but from a social standpoint, blogs have been described both as a primarily interpretive genre of writing “link-centered, highly interconnected filters of web content” (Herring et. al. 2004, 2005), and more broadly as a *n-to-?* medium in which a known number of authors asynchronously shares content with a conceptualized audience of an unknown number of readers, where the majority of bloggers produce content for very few readers (Boyd 2006). Bloggers produce content for a wide variety of motivations, ranging from the creative to the journalistic, the personal to the political (Nardi et. al. 2004b).

Bolstered by relatively easy access to an enormous number of publicly available texts through blogs, researchers have examined the phenomenology of blogging—why people blog (Nardi et. al. 2004b), the various ways in which it is technologically supported (Du and Wagner 2006), and the configurations and practices of those who use blogs (Nardi et. al. 2004a, Zhao and Rosson 2009).

Blogs and other social media have made a splash in the arena of American politics. The year 2008 saw Barack Obama elected to the presidency, and his campaign’s use of digital media was repeatedly cited as one of his keys to victory, so much so that the L.A. Times dubbed him the “Social Media President” (Sarno 2009). Shortly after the election, the New York Times credited Obama’s “networked, open-source campaign” and a “self-publishing, self-organizing democracy” that it helped create (Carr 2008). The President’s mastery of social media was again cited in 2012 in the popular press as one of his advantages over competitor Romney (Rosenstiel and Mitchell 2012). The veracity of claims such as these have yet to be fully investigated, but it is undeniable that the proliferation of these internet-based technologies has had an indelible impact on American politics.

A number of extant theories of media and ICTs are relevant to understanding the role of social media in political action. First, there is a long tradition of research on the role of media, particularly news media in politics—it has been firmly established, especially in recent decades, that media are not merely reflective of political realities, but also exert a significant measure of external influence over political systems.

With regards to social media, there is a significant body of research in the discipline of Computer-Supported Cooperative Work (CSCW) that deals with the ability of social media applications to be used for coordination and collaboration (cf Begel et. al. 2010, Muller et. al. 2012). The intersection of these areas, however, remains incompletely explored, though it is certainly a topic that has garnered some interest among serious academics. Some have theorized that social media's potential for political disruption lies in its facilitation of easy, decentralized information sharing (cf Benkler 2006), and some studies have been conducted that seem to corroborate that such information sharing does take place (Lotan et. al. 2011). That said, academic accounts of social media use in political action are incomplete and—at times—inconsistent. Even the political efficacy of social media use is disputed, with antipodean pronouncements ranging from the skeptical to the technoutopian having been made by prominent figures such as Castells (1996, 2011), Shirky (2003, 2011) and Gladwell (2010, 2011). It cannot be overstated that there is still much to be learned about the shape and content of political use of social media technologies, and that to do so necessitates the use of empirical methods.

With the plethora of types of social media available for study, each of which is shaped by different technical capabilities and is situated in different social contexts, a detailed study may benefit from focusing on a particular medium. Blogs have received particular attention as a medium through which many users have successfully self-published on the web, and the medium has gained some notoriety over the past decade for having influenced traditional news agendas. Over the past decade, a number of American political fortunes have risen and fallen due to public attention generated by blogs, perhaps most notably exemplified in the case of the resignation of Senate Majority Leader Trent Lott after widespread condemnation of controversial comments made at Strom Thurmond’s birthday (Drezner and Farrell 2004). Traditional media had largely ignored the comments but persistent publication by political blogs made the gaffe impossible to ignore (Shachtman 2002).

Over the past decade, political blogs have come into their own. A 2004 New York Times piece wryly commented that despite their obvious influence, “never have so many people written so much to be read by so few” (Hafner 2004). Since then, a few political blogs have become frequently linked in mainstream media sources and have accrued a number of the professional characteristics of mainstream news organizations. It is no longer uncommon, for example, to see bloggers with press credentials.

In the CSCW and communications literature, blogs have been largely addressed as a social medium—a form of mass publication and a venue for mass participation in the political process. While it is clear that blogs have an important role in politics, it is no longer clear whether all blogs may be uncritically considered to constitute the sort of massive citizen participation that has been the popular characterization of social media in the political process. It is, however, not a foregone conclusion that professionalized blogs can simply be modeled using our understanding of traditional media—while their content is made available online much in the same way as the internet presence of traditional news outlets, blogging organizations do not have the same structures and institutional concerns as traditional news organizations. Certainly, political blogs remain an interesting and important area of inquiry, and it is clear that the vast majority of blogs have not been professionalized. It would behoove us to learn how to think about the influence of these blogs in the political field.

A number of attempts have been made to describe the linkages between social media and mass media. For example, Scott (2004) examines the critical fact-checking and agenda-setting role of blogs in the popularization and discussion of Senator Trent Lott's controversial comments at Strom Thurmond's birthday party, finding that blogs were influential in calling and maintaining attention to the story. Many, including Lotan et. al. (2011) and Golan (2014) have described a symbiotic relationship between mainstream media outlets and individuals as exercised through social media. These efforts have not been limited to ICT researchers—political scientists (cf Drezner et. al. 2004) have argued that blogs have successfully constructed interpretive frames that act as focal points for mainstream media.

Attempts have also been made to understand the structure and content of politically oriented social media, with blogs and micro-blogs taking a front seat in much of the research. Computational efforts, such as that by Williams and Gulati (2008) found that social media content can act as an indicator for political preference, further corroborated by studies (e.g. Tumasjan et. al. 2010) that have found reflections of political realities such as polling numbers through analysis of microblogging data. There have been investigations of the structure of the political blogosphere (and Twitterverse) using network analysis (e.g. Adamic and Glance, 2005; Livne et. al. 2011) to identify communities within political bloggers[1]. These investigations into structure have repeatedly found that the American political blogosphere is bifurcated—liberal blog authors mostly link to one another and liberal commentators mostly comment on liberal blogs, with the same being true for conservative bloggers and commentators. This bifurcation, while interesting in and of itself, presents some opportunities for research as well—while American political blogs share the same subject matter, they do so in distinct, easily identifiable communities. This invites comparisons not just of liberal and conservative bloggers, but perhaps the ideologies of liberalism and conservatism themselves, as expressed through writing.

[1] See pgs. 27-28 for more on computational efforts.

In summary, what these studies make clear is that hidden in the massive amounts of text recorded in these digital media is a vast wealth of information about their users, their audiences, and their place in the information ecosystem.

## 1.3 Political Ideology

### *1.3.1 Ideology*

In political science, the concept of *political ideology* has been plagued by what one academic famously termed “semantic promiscuity,” (Gerring 1997) with this difficult and “elusive” concept approached from many different theoretical and analytical perspectives. Despite (or possibly due to) the age of this modern concept, which its progenitor Count Antoine Desutt de Tracy first described in his 1817 publication *Elements d’Ideologie*, disagreements continue as to the scope and precise meaning of the term (Kennedy 1979).

A plethora of definitions for ideology are still invoked in contemporary political science literature, with the scope and orientation of the definition provided heavily dependent on the research goals of its proponent. As the purpose of this dissertation is not, strictly speaking, to resolve disciplinary semantic disputes, for the purposes of our study we chose a “neutral,” textbook definition for ideology: “a set of beliefs about the proper order of society and how it can be achieved.” (Erikson and Tedin 2003; Jost et. al. 2009).

Generally, definitions of ideology agree that the beliefs that make up an ideology have a social element, in that they involve “shared frameworks” that are possessed by social groups or collectivities (Parsons 1951). Though they are, in essence, simply collections of beliefs, to evaluate them solely as such would be fairly identified as excessively reductionist. Ideologies are coherent structures that are characterized by a degree of *stability*, which can aid in the communication of the beliefs of identifiable constituencies. In addition, they are characterized by presentation in *contrast* to opposing ideologies (Knight 2006).

The formation and transmission of ideology has been a subject of intense curiosity, having been explored by early social theorists, and later, social psychologists, political psychologists, sociologists, economists, and cognitive psychologists. We find Weber’s perspective particularly insightful—he presents the notion of a “selective process” wherein people and ideas choose one another. This idea of reciprocity acknowledges the reality that individual agency is bounded by environmental factors, and this understanding conforms well to modern understandings of ideology. Generally, it is accepted that “ideological outcomes result from a combination of top-down socialization processes and bottom-up psychological predispositions.” (Jost, Federico, Napier 2009).

### *1.3.2 Political Partisanship in the United States*

We turn now to the most salient collective expression of political ideology in the United States—its political parties. Though party and ideology are not synonymous, the

Republicans and the Democrats—the two major parties in United States politics—both have a popularly recognizable ideological core. Republicans promulgate a conservative ideology, Democrats a liberal ideology.

As this dissertation is largely focused on linguistic differences found across partisan orientations, it is important that we define the terms “liberal and “conservative” as we will use them in this dissertation, since these terms have multiple meanings which differ by context and which encompass a number of ideological positions. Even when considering the United States alone, the terms “liberal” and “conservative” have shifted in meaning a number of times over the course of history.

While conservatism as we understand it today has existed for nearly a century, the term has only been used unqualified in reference to it since the 1950s. The modern conservatism to which we refer is that which responded to the various political and social upheavals of the latter half of the 20th century. These included the Cold War, the Civil Rights Movement, the 60s counterculture movement, and the deregulation of the economy in the 70s and 80s. It is characterized by support for republicanism, respect for tradition, “the rule of law and the Christian religion,” and a defense of “Western Civilization from the challenges of modernist culture and totalitarian governments” (Schneider 2009). Within the umbrella of conservatism, distinct ideologies can be identified. Fiscal conservatives support small government, free enterprise, limited regulation, and low taxes. Social conservatives defend traditional social norms and Judeo-Christian values, and tend towards patriotism and nationalism while opposing multiculturalism and immigration (Beer et. al. 2014). Other factions associated with conservatism include libertarians, neoconservatives, and paleoconservatives. For the purposes of this dissertation, we will refer to all of these ideologies collectively as conservative.

Liberalism in the United States has been similarly shaped by the historical events of the 20th century, existing in its current iteration, referred to by historians as “modern liberalism,” since at least the 1930s. Exemplified by President Roosevelt’s New Deal and President Johnson’s Great Society, it is a philosophy centered on the “unalienable rights of the individual.” These include freedom of speech, freedom of the press, freedom of religion, separation of church and state, right to due process and equality under the law. In the United States, it is characterized by support issues such as voting rights, civil rights, environmental justice, and social services like health care and education (Jeffries 1990)

### *1.3.3 Political Polarization in the United States*

America is obsessed with differences between liberals and conservatives, with an almost tribal obsession placed on political beliefs. Polarization is the separation of politics into partisan camps, and in American politics it has grown tremendously. A Pew Research Center (Mitchell et. al. 2014) report from 2014 argues that partisan polarization has become much more pronounced than even just one decade ago. Pew used a wide variety of measures to support this claim. One example is that according to the report, 92% of Republicans are to the right of the median Democrat, and 94% of Democrats are to the left of the median Republican. Another example highlights the skyrocketing level of partisan animosity—nearly a third of politically affiliated Americans now see the opposing party as a threat to the well-being of the nation itself, compared with less than a tenth two decades ago. These findings have been corroborated by contemporary research, with findings based on political blogs (Lawrence et. al. 2010), and even Twitter (Conover et. al. 2011).

Polarization is also affecting political outcomes, with commentators comparing it to a team sport: "Party identification predicts the vote because partisans pull for their team and the social groups that it symbolizes while at the same time rooting against the other party and its allied social groups" (Green et. al. 2002). Conservative and liberal have increasingly become synonymous with Republican and Democrat. Moderates are vanishing from Congress. However, the nature of this polarization is a matter of debate: some suggest that polarization in politics exists but it is driven by political elites, not the masses (McCarty et. al. 2006). However, others argue that the positions and priorities of parties have pulled apart, becoming increasingly defined around a set of popularly shared core beliefs (Iyengar 2005, Baldassari and Gelman 2008).

Whatever the reason for polarization, it is not without consequence: as more Americans hold polarized beliefs, people with polarized beliefs have become less willing to compromise or engage in interpersonal relationships with those with differing viewpoints Also, research has suggested that political polarization can lead to the adoption of more inefficient and inffective policies (Schulz 1996). The news media has become increasingly partisan, and researchers have argued that partisan news outlets engage in editorial filtering in which judgments of newsworthiness are intertwined with validation of accepted political narratives, impairing access for consumers of news to potentially valuable information (Baum and Groeling 2008).

Given the potential deleterious effects of political polarization, and also given that it seems to be a dominant narrative in the contemporary political narrative of the United States, it is little wonder that this is a topic of much interest in the research community. Perhaps in parallel with the polarization itself, more scientific literature is supporting the idea that there are fundamental differences between liberals and conservatives that extend beyond their behavior at the polls or their demography. Numerous psychological studies have been conducted, and theories have been posited to explain these differences (Jost et. al. 2008; Jost et. al. 2003, Graham et. al. 2009, Conover and Feldman 1981; Levenson and Miller 1976). Some studies have even used fMRI to determine if evidence exists for neurological differences between people of different political affiliations, largely finding identifiably different patterns of neural activation in people of differing political affiliation (Zamboni et. al. 2009). Naturally, these differences have implications for social science research—a recent study reported in *Science* revealed that conservatives report greater happiness while liberals display more evidence of happiness-related behaviors (Wojcik et. al. 2015). Are liberals and conservatives so innately and fundamentally different? A growing chorus of scientists suggests that it is quite possible that this is the case.

### *1.3.4 Public Choice Theory*

For many decades, political scientists modeled liberals and conservatives as economic actors, with individuals voting for politicians who would pursue policy decisions benefitting them, referred to as *rational choice* or *public choice* (Dunleavy 2014). Rational choice models were used to explain many phenomena in political science, including the function of bureaucracy, the role and effects of money in politics, the influence of interest groups, and voting patterns among various demographics (Dunleavy 1992). The economic models this tradition employed were frequently better at explaining and predicting reality than the normative and behavioral accounts that characterized earlier incarnations of political science.

However, since these theories were first posited, a healthy skepticism has begun to emerge. An example of an area in which this skepticism surfaced is in the discussion of when and how voting decisions are made in elections. The traditional, public choice school of thought suggests that candidates think about elections from an issues perspective, and that as voters gain information about where the candidate stands on issues they make up their minds about which candidate to support.

Another school of thought that suggests that the way the electorate is swayed is different, that people are not issues voters to the degree that was previously suspected, and that certain issues can dramatically come to the forefront in a distinct way, swaying voters both on their opinion on the issue and on the candidate. Shaw (1999), for example, writes that certain campaign events, especially national conventions, have relatively durable effects on elections, including galvanizing voter support for candidates and against the opposition candidate. Collingwood et. al. (2012) write that notification of primary results from states can inform voters and provide legitimacy to front-runner candidates. Benoit et. al. (2012) write that debates have effects that lead voters to support certain candidates more or less—but that the effects they have are not simply explained by increasing issue knowledge. They suggest that the debates can change *which* issues voters consider salient. Interestingly, the debates do not usually change voters' perceptions of candidate competence, but there are changes in perceptions of character.

These findings cannot easily be accounted for by the utility-maximizing economic principles employed by public choice theorists (cf Hamlin and Jennings 2011), and instead speak to the surprisingly central role of narrative, values, and complex, difficult to pin down concepts like legitimacy.

Gore Vidal has famously observed that "the genius of our system is that ordinary people go out and vote against their interests" (Dreifus 1986). Clearly, this message resonated with many, because academics began to speculate: why was the rhetoric surrounding campaigns often about issues that seemed to have little economic impact, and why was that rhetoric often more charged than the issues that would have the most effect on the livelihood of voters? A chorus of doubts emerged as to whether positivist rationality truly governed behavior at the polls. Many questions have been raised about the assumption of rationality by pointing out the failure of rational choice models to account for concepts such as altruism, and by highlighting errors these models commonly made in conceptualizing the “nature of human goals” and “the processes that people use in reasoning from their actions to their values” (Simon 1995).

To this day, rational choice theory remains to some extent central to economic and political science research. However, it is much more carefully applied, and it is now recognized as one of multiple approaches that can be taken to understand different slices of political phenomena.

### *1.3.5 Moral Foundations Model*

With the preeminence of the rational choice models in serious question, many academics pursued the development of alternative frameworks with which to explain the aspects of human behavior they were observing. Abandoning the utilitarian and rationalist approach, a number of researchers would claim that at the core of politics is something more fundamental than even the maximization of personal benefit—ethics and morals.

Haidt (2007), in an attempt to explain cultural differences, posits five *moral foundations*, which are psychological systems that elicit strong emotional reactions: harm, reciprocity, ingroup, hierarchy, and purity. The applicability of these moral foundations to understanding ideology and politics was quickly apparent. In reaction to numerous reports of seemingly counterintuitive voting behavior among liberals and conservatives (cf Sears and Funk 1991, Berinsky 1999, Kinder 1998, Miller 1999), an increasing number of studies now support the thesis that liberals and conservatives have entirely different moral foundations (Haidt et. al. 2009, Graham et. al. 2009). It is speculated that a difference in moral intuitions between liberals and conservatives might mean that while both liberals and conservatives are deeply morally driven in their formation of policy positions, when analyzing conservative agendas “liberals may not recognize [them] as moral at all,” since conservative morality is dependent on dimensions not at all found in liberal psychology: ingroup, hierarchy, and purity. For liberals, only harm and reciprocity are seen as important (Haidt and Graham 2007).

While the taxonomy described above has been well-received and influential in academic circles over the past several years, the authors themselves note that it is not the only possible configuration of values, virtues, or vices that could underlie political opinion formation and voting behavior—indeed, others (cf Rai and Fiske 2011) have proposed other moral foundations for use in other contexts. Cognitive linguist George Lakoff frames political choice as contextualized by morally-grounded folk psychological metaphors—the “nurturant parent” versus “strict father” frames of mind (2010).

Critically, that the moral conceptualization of politics has come into vogue in recent years has meant that there now exist multiple prominent and widely accepted psychological models for political choice. On one hand, rational choice models remain useful and popular for certain types of analyses. On the other hand, the clear validity of alternative frameworks that rely not at all on the psychological factors underlying personal utility maximization have made it clear that we do not have a complete model of the psychology of political choice. Few attempts have been made at reconciling public choice and moral foundations, and it is unclear whether such reconciliation would be scientific or metaphysical—every public choice, after all, is “based, implicitly or explicitly, on normative premises that are crucial to its practical effectiveness.”

Those attempting to find theoretical grounding for any explanation of empirical results are faced with many competing theories and a dearth of definitive answers. Some theories, like Lakoff’s, have attracted significant attention but have lacked empirical support. A key component of the reflexive process of generating theory for the psychological bases of political partisanship is the pursuit of empirical evidence. It is without question that theoretical frameworks must be carefully examined in the context of real-world evidence.

# Chapter 2 : Textual Analysis of Political Writings

## 2.1 Background

In the previous chapter, we motivate a discussion of American political ideology, providing necessary background for our study of blogs. In this chapter we discuss approaches to the study of text, paying particular attention to social media texts, blog texts, political texts, and the intersection thereof. Each of the aforementioned have been the subject of considerable attention from multiple disciplinary perspectives.

Content analysis is a term referring to the various research techniques that focus on the "objective, systematic, and quantitative description of the… content of communication" (Berelson 1952, Kassarjian 1977). At its essence, it focuses on the processing of information in which "communications content is transformed, through objective and systematic application of categorization rules, into data that can be summarized and compared." (Paisley 1969). Kerlinger (1964) notes that content analysis, in addition to being a method of analysis, is also a method of *observation*: "Instead of observing people's behavior directly, or asking them to respond to scales, or interviewing them, the investigator takes the communications that people have produced and asks questions of the communications."

In content analysis, units of measurement are determined—a variety of categories textual elements were commonly quantified. For example, Flesch (1951) in his famous analyses of readability, focused on counts of categories of words. Others focused on counts of higher-order elements, such as theme (Holsti 1968), characters in literature, and space-time measures (Kassarjian 1977).

Next, categories within these selected elements were carefully classified: for example, in categorizing communications using words as the unit of analysis, Flesch coded words as compound or simple, as being part of a prepositional phrase or not, as being difficult or easy. Spiegelman et. al. categorized comic characters by their race or species, and categorized scenes by their locations (1952).

Early on, content analysis was employed mainly for analyzing media content—changes in newspapers and magazines over time, and across various works of literature. Such studies were largely descriptive; two prominent examples deal with the linguistic characteristics of newsworthiness (Galtung and Ruge 1965), and changes in public attitudes about issues such as race and ethnicity (Barcus 1961). These methods were sometimes applied to the study of politics—a notable example being that the speeches of Soviet Politburo members were analyzed for insights into the otherwise opaque internal power dynamics of the USSR (Hermann 1980).

Content analysis was popularly utilized by communications researchers well before the advent of computer-based research methods, and researchers noted the difficulty of quantifying large amounts of text, often referring to “the immense task of analyzing existing documents” (Kassarjian 1977). Sampling was widely and necessarily used. The usefulness of computers was quickly recognized, though, and as early as the 1960s, researchers began to develop analysis tools for natural text, most notably the General Inquirer System (Stone et. al. 1966). The General Inquirer could “locate, count, and tabulate text characteristics,” in which words that were coded as belonging to one or more categories (as specified in dictionaries) were counted, as it had been done manually by earlier content analysts. Initially, text needed to be punched onto computer cards, but soon, more advanced digital input methods and user interfaces made analysis easier.

Tools like the General Inquirer System illustrated the power of content analysis—by using or creating dictionaries of words experimentally confirmed to have positive or negative emotional associations, researchers could create a measure for the polarity (positivity or negativity) of a given document in a manner that could be automated by computers (Yi et. al. 2003 “sentiment analyzer”).

### *2.1.1 Digital social research*

Much has changed since the advent of content analysis. Though we likely owe much to the discipline of content analysis for our approach to natural language texts today, far from all quantitative and social studies of texts today explicitly identify as being part of the tradition of content analysis. We speculate that this is because academics from many disciplines, enabled by ever-improving digital tools, have naturally come to find digital analysis of texts as a useful direction for their research. Furthermore, the explosion of social activity over the Internet and its important place in everyday life has meant that researchers of human behavior are often now *necessarily* conducting observations of behavior mediated by computers, observations of which are also often text-based and computer-mediated. For example, CSCW emerged with the goal of supporting and understanding computer-mediated activity, furiously borrowing methods from a multitude of disciplines (often including content analysis, cf Newman et. al. 1995) with an ultimate goal of systematically integrating the process of understanding and design in this problem space (Ackerman, 2000).

With considerable academic attention focused on digital and digitalized texts, many new methods for analysis of natural language texts have been developed in the past several decades, generally referred to as “natural language processing” (NLP). For example, sophisticated machine learning algorithms have been developed to analyze text. Algorithms for topic modeling, for example, have enabled computers to discover “topics,” statistical co-occurrences of words that are “discovered” as a latent property of the texts being analyzed. By assuming that written documents contain a number of such “topics,” and that these topics will be distributed across multiple documents, algorithms produce groupings of words that are then identified by researchers as “topics” in the abstract linguistic sense. (Blei et. al. 2003, Blei and Lafferty 2009).

Topic modeling is just one example of a natural language processing technique. Extracting semantic data from text is a difficult task due to the multiple layers of abstraction involved in natural langauges. Though it is relatively simple to extract words from text, *stemming* words—removing morphological elements from the root of a word—is subject to numerous, complex rules. Even more complicated is the parsing of phrasal expressions (e.g. “find out,” “take up”), especially because parts of phrases are separated by other words depending on context: “I found it out the next day.” Finding the meaning of entire sentences adds several more layers of complexity, since grammar is ambiguous and subject to context and interpretation. The four-word sentence, “We saw her duck” could be interpreted in a handful of ways, depending on which definitions and parts of speech the component words take. If we consider just two of the possible meanings of the sentence, we could have noticed a woman’s pet, or we could have observed her lower her head or body quickly. Such ambiguity is the rule, not the exception. However, in everyday use, meaning is usually apparent by some tacit understanding of the context, as is the case with these fictitious headlines: “Kids make nutritious snacks,” and “Grandmother of eight makes hole in one.”

The sheer vastness of the data now being collected, stored, and potentially made available for study has been widely noted; A McKinsey study in 2011 (Manyika et. al. 2011) estimated that at the time of publication, 7 exabytes of consumer data was being stored by enterprises and 6 exabytes were stored on personal hard drives. According to that study, one exabyte of data “is equivalent of more than 4,000 times the information stored in the US Library of Congress.” Obviously, not all of this data is available to researchers, but the amount of data analyzed in individual social science studies has been growing in recent years. Whereas early content analysis papers examined a few documents at most, today, thousands or even millions (cf Vieweg et. al. 2010) of documents are analyzed at a time.

The advent of big data in social research has drawn commentary that the empiricism of such ‘big data’ research is often exploratory and descriptive, and is still “characterized by a focus on the size of the data set”. Some have decried this vein of research as spurious pattern-finding (Marres and Weltevrede 2012). This is unsurprising given the access to categories of research questions newly enabled by big data, for example observations of written data across a multitude of subjects.

It is clear that there is much opportunity to be found in analyzing large corpuses of text, as even older textual analysis methods may be used to great effect on populations that it was previously difficult to study. For example, though the aforementioned use of the General Inquirer system for sentiment analysis emerged relatively recently, dictionary-based methods have been in use for decades. Though the basic technology behind these methods is simple (classify words into meaningful categories, and count the prevalence of these words in documents), both the classification methodologies and automation technology have been improved. A prime example of this is the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker 1993, Pennebaker et. al. 2007, 2011), a dictionary-based content analysis tool that counts function words and words related to several different classes of psychological processes. We discuss this tool in more detail in the following chapters.

### *2.1.2 Function Words*

Though this dissertation examines the use of many of the categories analyzed by LIWC, we focus in particular on the use of *function words*. While much attention in textual analysis has been focused on content, a relatively newer area of inquiry deals with the structural side of language. Function words—the short, forgettable words that have little lexical content but establish grammatical relationships with other words, have been revealed to be revelatory of personality, demographic, and other social metrics. Function words make up the structure of language, consisting of categories of words like pronouns, articles, and conjunctions.

When researchers shifted their attention from the lexical content of texts to examine the use of these forgettable words, their use was found to be correlated with factors such as sex, age, power, truthfulness, interpersonal relationships, and even love (Phillips 1973, Tausczik and Pennebaker 2010; Penenbaker et. al. 2003; Koppel et al. 2002). Pennebaker, writing on the surprisingly revelatory nature of these words, calls them "the keys to the soul."

People who use high rates of personal pronouns—including I, we, you, she, and they—tend to be self-reflective and highly social (Chung and Pennebaker 2007, Pennebaker 2011). I-words, specifically, track where people are paying attention. If people are self-focused, insecure, or self-effacing, they tend to use first person singular pronouns at high rates. If confident, focused on a task of some kind, or lying, their rates of using I-words drop. In one study of blog posts written immediately before and after the 9/11 attacks, a sharp drop is observed in the use of "I" and a sharp increase the use of "we" among American bloggers as well as then-President Bush, demonstrating that a sense of togetherness and national identity manifested in the speech of Americans following a deeply-felt national crisis (Cohn et. al. 2004). Pronoun use as a result of a community-wide tragedy also responded similarly (Gortner and Pennebaker 2003). In the same vein, another study (Mark et. al. 2012) found that use of “I-words” decreased relative to the use of “we-words” as violence during the Iraq war increased. It has been proposed that first person pronouns demonstrate a degree of focus on the self: they are used at higher rates among people who are depressed (Rude et. al. 2004), suicide prone (Stirman and Pennebaker 2001), honest (Newman et. al. 2003), or lower in social hierarchies (Kacewicz et. al. 2013).

While many fascinating studies have revealed links between differences in pronoun use and behavior, other categories of function words have also been examined. The frequency of use of articles, verbs, conjunctions, negations, prepositions, and other types of words were found to be predictive of social dynamics in an experimental setting (Gonzales et. al. 2009). Examination of documents in the field have also found that there is a significant difference in the way people of different genders use many categories of function words (Newman et. al. 2008).

### *2.1.3 Politics and Textual Analysis*

Unsurprisingly, computerized textual analysis has been employed in contemporary research of politics. A number of comparative analyses have been performed on political speeches, finding patterns and trends across winners and losers, across political lines, or over the course of a political career (Weintraub 1986, Chung and Park 2010, Slatcher et.

al. 2007). Increasingly sophisticated computerized analyses have been identified as useful in the prosecution of this line of research (Lucas et. al. 2013)

Prediction is a common theme in textual analysis research. Efforts have been made to automatically classify documents by ideological affiliation (Koppel et. al. 2009, Conover et. al. 2011). Some have tried to forecast elections by mining Twitter text data (Tumasjan et. al. 2010, Sang and Bos 2012), and, unsurprisingly, stock markets (Zhang et. al. 2011). We note, however, that a deep undercurrent of skepticism abounds regarding this trend in research, alleging that critical errors and faulty assumptions are often made (Gayo-Avello 2012). Almost all of the above studies have taken place with data from blogs or microblogging platforms. In addition to being relatively easy to collect, as the subject of research, blogs have the advantage of being written by a diverse population on a diverse set of subjects, political or otherwise.

In addition to finding correlations between textual data and real-world outcomes, textual analysis has been employed to examine the relationship between the content of texts and observable behavior. Some studies examined the values and motivations of online contributors (Chen et. al. 2014), the political influence of individuals in online communities (Stieglitz and Dang-Xuan 2012), and changing morality in politics (Motyl 2012). Another experimental study found that when subjects were primed with moral language, such as an invocation of the need to care for children or the need to preserve a way of life, liberals and conservatives were more likely to show evidence of entrenchment in their existing political attitudes (Day et. al. 2014).

As previously established, liberal and conservative blogs have been shown to belong to separate communities that only interact in limited ways: Almquist and Butts (2011) used a novel topic modeling algorithm to demonstrate the insularity of liberal and conservative citation patterns in blogs, and previously, Adamic and Glance (2005) used network analysis to show the divide in the liberal and conservative blog populations. Though some exploratory studies (cf Yarkoni 2010) have been conducted on textual markers in blog language, little exploration has yet been conducted on the political blogosphere.

The current understanding of linguistic anthropology (cf Gumperz 1964, Labov 1972) holds that insular groups can develop different linguistic characteristics as markers of membership. As such, we wonder whether this increasingly bifurcated group of bloggers may be developing different verbal characteristics.

## 2.2 Research Questions

As discussed earlier, in the past decade, there has been a resurgence of investigations in the social sciences that focus on the analysis of natural language. In the past, "the analysis of text has been slow, complex, and costly" (Chung and Pennebaker 2007). However, much has changed in recent years. First, the availability of natural language texts on the internet has expanded greatly. In addition, the development of new computer text analysis methods (cf Pennebaker et. al. 2001) means we have the ability to examine social processes in novel ways. Research using quantitative analysis methods on digital texts has revealed that what were previously considered to be "junk words" are laden with meaning and are associated with a wide variety of social behaviors (Chung and Pennebaker 2012).

We believe function words to be like textual fingerprints—just as one might unthinkingly leave fingerprints, the use of function words is unavoidable and largely unconscious. Though easy to ignore, examination can reveal much about the behavior of those that use them. Function words are keenly tied to emotion: one of the first systematic investigations involving a large spectrum of function words looked at the language used by populations of students suffering from depression and bipolar disorder. By comparing the language of a control group with language used by depressed students, it was found that depressed students used first person singular pronouns more relative to non-depressed students (Rude et. al. 2004).

At the same time, an increasing number of academics studying ideology have begun to find support for the claim that the political stances that make up ideology are more than simply utility-maximizing determinations made on individual bases—we believe ideology to be systems of thought that are shared and transmitted between individuals and communities (Thompson 2013), inculcated in a foundation of deep-rooted moral foundations (Graham et. al. 2009).

The factors affecting various types of function word use are myriad—much as with fingerprints, it is difficult to form a holistic narrative that explains all the variation found in function word use. However, a number of recent successes in detecting moral rhetoric in texts through correlation function word counts and other dictionary-based word counts suggests that an individual's moral underpinnings have systematic effects on the text that they produce (Graham et. al. 2009, Sagi and Dehghani 2014, Day et. al. 2014, Vaisey and Miles 2014, Motyl 2012).

While we are careful to recognize the limitations in our methods, we find value in exploratory, bottom-up research, as many of the basic questions involving these types of linguistic markers have not yet been answered. This bring us to our primary research question:

> Q: Do there exist meaningful differences in liberal and conservative use of linguistic markers such as function words in their writing, and if so, what might explain those differences?

This question can be broken down into a number of subordinate inquiries: In terms of mean use of function words and psychological process words as measurable in LIWC, are there significant differences (beyond natural variance) between the textual output of the liberal and conservative blogging populations? Do differences exist in liberal and conservative use of these linguistic markers when aggregated by date?

We also seek to explore the implications of any findings that come about as a result of this inquiry: What might be the psychological mechanism causing differences and similarities in function word use between these two populations? If no significant differences are observed in function word use between liberals and conservatives, we might find less support for claims that fundamental psychological factors like morality are driving the way liberals and conservatives talk about politics. Do liberal and conservative language use patterns correspond to previously identified patterns in similar research across different populations? If so, we might be led to believe that similar psychological processes are at work in shaping ideological language. Additionally, a natural question that arises in attempting to answer these questions is how one might control for other factors governing word usage, such as population characteristics.

Finally, in examining how linguistic marker use changes in liberal and conservative blogging populations when aggregated by date, we prosecute a use of textual analysis that few to date have attempted. The challenges of comparing linguistic markers on a day-by-day basis are clear: from document to document there is a huge amount of contextually-driven variation in language use, so a relatively large amount of data must be collected. The American political blogosphere is an ideal population for this novel use of textual analysis, because we may observe two demonstrably separate communities that are nonetheless thematically linked, with both liberal and conservative political blogs being tied together by a common subject matter. Furthermore, American political blogs are on a short-term basis *topically* driven in significant ways by the American news media, in part mitigating one possible source of linguistic variance (Cornfield et. al. 2005). The contemporary understanding of use of linguistic markers is faced with a seemingly dissonant set of claims: first, that their use by individuals remains relatively invariant even over a period of years (Holmes 1998), but that inter-document variation even within individuals is extremely high (Chung and Pennebaker 2007). It is our hope that this study can help unpack this contradiction, improving the scholarly understanding of the study of language.

## 2.3 Dissertation Outline

Chapter 3 will describe the methods we have undertaken to answer our research questions, which will include discussion of the data set and the process of collecting it. We will discuss the methodological considerations of collecting writings publicly available on the Internet, and of analysis of their linguistic features through computerized tools.

In Chapter 4, we discuss theories that underlie conservative and liberal political stances, focusing on the popular conceptualization of the psychological underpinnings of liberalism and conservatism in America pioneered by George Lakoff. We show how we can use linguistic tools to test Lakoff's thesis that liberal and conservative philosophies are fundamentally gendered, by examining whether linguistic features of partisan blogs reflect established patterns in gendered writing. In finding that liberal and conservative blogs do exhibit patterns of language use that closely parallel those patterns found in language use by men and women, we find empirical support for Lakoff's model of political psychology.

Chapter 5 builds on the findings in Chapter 4, further identifying differences in liberal and conservative writings. In this chapter, we focus on the concept of *linguistic style*, broadly defined as the phonological, lexical, syntactic, prosodic, and orthographic variation observed within a single language, as utilized by individuals and groups. We find evidence of lexical and syntactic differences between liberal and conservative language use and discuss the implications this may have for our understanding of ideology and political polarization.

Chapter 6 examines the use of function words over time. We investigate whether the short-term fluctuations in use of function words are noise, or if they reflect a response to some outside stimulus. We find that the use of function words over time by the liberal and conservative populations are significantly cross-correlated. As this finding is suggestive of some underlying condition or stimulus to which liberals and conservatives language is responding in the same way, we investigate what the mechanisms underlying this observed behavior may be.

# Chapter 3 : Methods

## 3.1 Background

Our overall methodological approach will involve the scraping of corpora of blog data centered around the 2012 United States presidential election. For analysis we employ a computational linguistic analysis tool, LIWC, to produce usable data from the large textual corpora. We examine linguistic markers, including the use of function words, for trends, internal correlations, and correlations with externally obtained data sets such as polling data, and characteristics that conform to real-world events. As the relationship between political ideology and function word remains relatively unexplored, we take a “bottom-up” exploratory approach to this study, for which these methods are particularly appropriate to tackling the research questions posed. As shown by the conflicting accounts of the influence and role of social media to this day, we continue to lack an understanding of the content of blogs and how that content relates to the external world. Any insights gleaned from an in-depth quantitative analysis of this textual data will contribute to the ability of researchers to theoretically situate blogs in the information ecosystem, regardless of what the particulars of the findings may be.

## 3.2 Data Collection

### *3.2.1 Coding*

We collected the top-ranked U.S. political blogs from Technorati, a blog search engine and directory that ranked blogs based on a proprietary authority measure. We then coded these blogs by their political ideological leaning and host name. As of the time of this writing, Technorati no longer publishes this formerly authoritative listing of blogs. Technorati's political blog listing was well-respected, and has been used previously in academic research (cf Adamic and Glance 2005)

To determine whether a blog was liberal or conservative, we had to manually examine each blog for evidence of political orientation. To do so, we used a multi-tiered process involving two coders, both of whom were politically engaged U.S. citizens, and one of whom was a political scientist:

1) Each coder independently checked for obvious, explicit markers of political affiliation. Political affiliation of blogs was often openly provided on an "about us" page, or a sidebar. Example: Redstate.com advertises itself in the "About Us" page as "the singular hub of *conservative* grassroots collaboration *on the right.*" The vast majority of the blogs we examined were able to be coded immediately using this step. We also verified whether the blogs were centered around U.S. politics in the same way.

2) For blogs where no political affiliation was provided, each coder looked for membership in blog rings, which are often explicitly politically affiliated, such as the Watcher's Council, a conservative blog ring, or the Progressive Women's blog ring, a left-leaning blog ring.

3) Where the above steps did not succeed or were not applicable, each coder read a selection of twenty randomly selected blog posts from the blog in question and coded those posts for openly declared political affiliation, or the coders looked for other markers of political affiliation such as well-defined stances on popular issues. Blogs that did not have a large majority of posts that were clearly liberal or conservative were coded as having no clear political affiliation.

The codes generated by both coders were compared. If both coders identified a blog as being liberal or conservative, it was labeled as such. If the codes for a given blog were not in agreement, the blog was skipped. The coders proceeded down the list of Technorati's top blogs until 50 liberal and 50 conservative blogs were identified by a consensus of the coders.

### *3.2.2 Scraping*

Scraping is an attempt to make textual data useful by processing an existing rendition of it, compartmentalizing, labeling, and sorting it as necessary. While a number of commercial and free scraping tools are available—Python has libraries dedicated to crawling and scraping textual patterns, which are well-suited to blogs—we found that the wide variety of blogging platforms and the different deployments thereof caused myriad issues when we attempted to crawl each one manually. Some blogs had countermeasures deployed against crawling and scraping that were difficult to circumvent, since crawling and/or scraping behavior can look similar to malicious actions such as Distributed

Denial-Of-Service (DDOS) attacks, or are indications that an entity may be attempting to misappropriate copyrighted data.

We used a crawler program, written in Java, to automatically collect posts from the Blogspot feeds interface. Blogspot is the domain name that hosts blogs that use Google's popular Blogger publishing platform. One of the earliest blogging platforms, Blogger helped publicize the format and remains popular to this day. Because it is a hosted service, and because the platform provides a fairly easy API for syndication, we were able to scrape sites hosted on Blogspot relatively easily. Unfortunately, less than a third of the blogs we scraped were from Blogspot.

In fact, many of the blogs we scraped were custom-built, particularly the most successful blogs like DailyKos, Huffington Post, and RedState. These sites were also the most likely to have technical countermeasures against scraping. As such, for these blogs and for other Wordpress blogs (which often had various customizations and flexible page layouts), we used Mozenda, a commercial Web data scraping software. Mozenda is a scraping platform which allows users to create "agents," scrapers that the user trains to work on specific websites given certain user-determined parameters. Each blog had to be trained separately, with the user programming the sequences of clicks to get to the relevant data and defining the parts of the blog page the agent would obtain data from. The technology would use fuzzy logic to determine, for example, where the date of the blog post was, and where the name of the author was.

We found this approach to be much easier than our original approach of manually coding scrapers for each website. However, there were still issues—some blogs were still protected against scraping, despite Mozenda's sophisticated circumvention of the countermeasures employed by most of the blogs we had trouble with. Fortunately, we only encountered two whose countermeasures Mozenda could not defeat, and their blog posts in this date range numbered in the tens and low hundreds, so we manually copied the text from those blogs into our dataset.

Also, Mozenda's scrapers were far from perfect. If blogs had dynamic page elements like ads displayed among the listings of blog posts, they would often interfere with Mozenda's ability to find the link to the next blog post, requiring manual intervention to restart. Oftentimes, Mozenda would encounter errors for no discernible reason. Because of these difficulties, blog post collection took place over several months.

### *3.2.3 Summary*

Overall, we were able to obtain data from 50 individual liberal blogs (45,172 posts) and 50 conservative blogs (39,838 posts). After removing posts with no content, and posts containing only a link or picture, we were left with 43,478 liberal posts and 38,171 conservative posts. Our blog post data ranges from May 1, 2012, to Dec. 31, 2012, which spans the entire general Presidential Election campaign waged between President Barack Obama and the Republican challenger, Mitt Romney. The election was held on Nov. 6th, 2012.

The output of the data we obtained was in the form of CSV files. For the Blogger blogs, all of the blog posts were outputted into one database that was exported to CSV, with fields for Blogger ID (a numeric index corresponding to the blog in question), Blog Title, Blog Date, Blog Author, and Blog Text. We also collected comments for the Blogger blogs, but these were not used in the analysis. For the blog posts collected through Mozenda, we had a CSV spreadsheet file for each blog, with fields for Blog Title, Blog Author, Blog Date, and Blog Text.

## 3.3 Data Processing

As LIWC necessarily processes documents as individual text files, we needed to convert the CSV file into individual text files. In earlier iterations of this study, we did this data processing within Microsoft Excel, using Visual Basic macros. We attempted to aggregate all of our CSV files and parse it through Excel to output each CSV file into a separate text file. This was a mistake, as Excel crashed within five minutes of opening up each of our large files. Excel handles large amounts of text and extremely large files very badly. We instead wrote a python script to process the CSV files, and put them into a format that could be analyzed by LIWC.

### *3.3.1 Challenges to data processing*

Methodologically, the largest challenges we faced had to do with data processing. Parsing large amounts of textual data can lead to many unexpected errors, and our study was no exception.  The first major set of issues had to do with textual encoding. Some websites apparently encoded in UTF-8, and others encoded in ASCII. The scrapers we used, which lacked the ability to intelligently detect and convert the encoding from websites, often generated output files that contained one or more different encodings, which made it difficult to do textual processing. At times, this would lead to our data being read as gibberish. At other times only the punctuation in blogs would be garbled. However, since our data was exported by our scrapers as CSV files, altering the encoding of a file could sometimes break the CSV format—one example of an issue was that our parser did not recognize Unicode commas as delimiters.

Incompatible standards were also an issue when working with the data in our CSV files. First, the syntax for interacting with the filesystem differs between Windows and Mac OS, as do default text encodings. In addition, CSV files use different conventions between Windows and Mac OS as well, which sometimes caused issues with our parser. In addition, the libraries in python and Java that we used to write CSV files dealt poorly with writing cells that contained blog text. This was because blog texts contain numerous whitespace characters, punctuation marks and escape characters that could cause parsers to misread a file. Essentially every blog post we scraped needed to be manually scanned for errors resulting from incorrect parsing or writing.

Even heterogeneity in non-technical standards gave us difficulty. Dates on different blogs were often formatted differently and in plain text rather than in numbers, so each blog we scraped needed to be inspected and, if necessary, the date would need to be manually converted to a reasonable standard DD/MM/YYYY format. Interventions like these were very time consuming. However, the analyses we performed would have been essentially impossible without this thorough process of cleaning and controlling the data.

To prepare our data set for analysis, we removed all posts consisting of fewer than ten words, all posts that seemed to be duplicates, entries that did not seem to be blog posts but were nevertheless mistakenly scraped, such as advertisements. We manually fixed the text of posts whose data was improperly identified by text processing tools as syntactic elements of CSV files, removing any offending characters.

## 3.4 Analysis

### *3.4.1 LIWC*

I used the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker et. al. 2001, 2007), which is a powerful textual analysis tool that incorporates measures of polarity but also incorporates various affective measures such as sadness or happiness and provides pronoun counts, which have been demonstrated to be correlated to a number of different psychological factors such as perceptions of power, in-group/out-group dynamics, values, and personality (Pennebaker 2011).

LIWC essentially operates as a word count analysis tool that checks input documents against several dictionaries that its authors have curated. In addition to checking word counts of words that are identified within these dictionaries, LIWC counts other metrics such as sentence length, use of punctuation, the count of words over 6 letters long, and the total word count of documents. Since its publication in 2007, it has been used and cited in numerous academic publications. Many of its dictionaries have been tested for external validity, such as the words associated with negative and positive emotions.

LIWC's dictionaries and measures fall into several major categories:

1) Linguistic processes, which include measures like word count, all of the function word measures, and dictionary words.

2) Psychological processes, which are broken up into social processes, affective processes, cognitive processes, perceptual processes, biological processes, and relativity, each of which have subcategories as well.

3) Personal concerns, such as work, achievement, leisure, and religion.

4) Spoken categories, including assent, nonfluencies, and filler words.

A complete listing of the LIWC categories, copied from its documentation materials, are viewable in Table 3.1  below.

| **Category** | **Examples** | **Words In Category** |
|---|---|---|
| Word count | | |
| words/sentence | | |
| Dictionary words | | |
| Words>6 letters | | |
| Total function words | | 464 |
| Total pronouns | I, them, itself | 116 |
| Personal pronouns | I, them, her | 70 |
| 1st pers singular | I, me, mine | 12 |
| 1st pers plural | We, us, our | 12 |
| 2nd person | You, your, thou | 20 |

|  |  |  |
|---|---|---|
| 3rd pers singular | She, her, him | 17 |
| 3rd pers plural | They, their, they’d | 10 |
| Impersonal pronouns | It, it’s, those | 46 |
| Articles | A, an, the | 3 |
| Common verbs | Walk, went, see | 383 |
| Auxiliary verbs | Am, will, have | 144 |
| Past tense | Went, ran, had | 145 |
| Present tense | Is, does, hear | 169 |
| Future tense | Will, gonna | 48 |
| Adverbs | Very, really, quickly | 69 |
| Prepositions | To, with, above | 60 |
| Conjunctions | And, but, whereas | 28 |
| Negations | No, not, never | 57 |
| Quantifiers | Few, many, much | 89 |
| Numbers | Second, thousand | 34 |
| Swear words | Damn, piss, fuck | 53 |
| Social processes | Mate, talk, they, child | 455 |
| Family | Daughter, husband, aunt | 64 |
| Friends | Buddy, friend, neighbor | 37 |
| Humans | Adult, baby, boy | 61 |
| Affective processes | Happy, cried, abandon | 915 |
| Positive emotion | Love, nice, sweet | 406 |
| Negative emotion | Hurt, ugly, nasty | 499 |
| Anxiety | Worried, fearful, nervous | 91 |
| Anger | Hate, kill, annoyed | 184 |
| Sadness | Crying, grief, sad | 101 |
| Cognitive processes | cause, know, ought | 730 |
| Insight | think, know, consider | 195 |
| Causation | because, effect, hence | 108 |
| Discrepancy | should, would, could | 76 |
| Tentative | maybe, perhaps, guess | 155 |
| Certainty | always, never | 83 |
| Inhibition | block, constrain, stop | 111 |
| Inclusive | And, with, include | 18 |
| Exclusive | But, without, exclude | 17 |
| Perceptual processes | Observing, heard, feeling | 273 |
| See | View, saw, seen | 72 |
| Hear | Listen, hearing | 51 |
| Feel | Feels, touch | 75 |
| Biological processes | Eat, blood, pain | 567 |
| Body | Cheek, hands, spit | 180 |
| Health | Clinic, flu, pill | 236 |
| Sexual | Horny, love, incest | 96 |
| Ingestion | Dish, eat, pizza | 111 |
| Relativity | Area, bend, exit, stop | 638 |
| Motion | Arrive, car, go | 168 |
| Space | Down, in, thin | 220 |
| Time | End, until, season | 239 |
| Work | Job, majors, xerox | 327 |
| Achievement | Earn, hero, win | 186 |

|  |  |  |
|---|---|---|
| Leisure | Cook, chat, movie | 229 |
| Home | Apartment, kitchen, family | 93 |
| Money | Audit, cash, owe | 173 |
| Religion | Altar, church, mosque | 159 |
| Death | Bury, coffin, kill | 62 |
| Assent | Agree, OK, yes | 30 |
| Nonfluencies | Er, hm, umm | 8 |
| Fillers | Blah, I mean, you know | 9 |

**Table 3.1. LIWC categories of words (Pennebaker et. al. 2007)**

LIWC's output is given in the form of a table where each row is a document and each column corresponds to one of its measures. For the purposes of this study, we used each blog post as the document unit. For each measure, LIWC either provides the calculated measure or the integer number of occurrences of words present in a given dictionary contained within a given document.

### *3.4.2 Generalized Linear Mixed Models*

One possible issue that we were acutely aware of over the course of our study was the possibility that random effects could provide confounding factors in our data. We also could not dismiss the possibility of collinearity between the measures provided by LIWC: in fact, as some measures were *subcategories* of others, it was absolutely certain that not all measures were not linearly dependent of one another, depending on which measures we decided to include into our model. As such, we could not simply do a test of means for each measure in which we were interested between our liberal and conservative data sets, nor would standard linear models suffice. In this case, the appropriate statistical tool was the Generalized Linear Mixed Model (GLMM).

GLMM is an extension to the generalized linear model where random effects in addition to fixed effects are accounted for, allowing some tolerance of non-independence among predictor measures. It follows the general form of:

$$y = X\beta + Z\gamma + \varepsilon$$

Where $y$ is an $n$-length vector of the outcome variable, $X$ is an $n$ $x$ $p$ sized matrix of $p$ predictor variables, $\beta$ is a vector of $p$ fixed effects regression coefficients, $Z$ is the $n$ $x$ $q$ design matrix for $q$ random effects, $\gamma$ is a vector of coefficients of the $q$ random effects, and $\varepsilon$ is an $n$-length vector of the residuals.

In all cases where we used GLMM in this dissertation, each blog post served as one entry and our target variable $y$ was the binomial variable indicating whether the post was coded as liberal or conservative. $n$ is the number of blog posts examined, which in our case was $n = 43{,}478 + 38{,}171 = 81{,}649$.

The predictor variables used were different combinations of LIWC measures depending on the goal of the analysis at hand, and the matrix $X$ would have one column each for each of the predictor variables, which might include measures such as the proportion of pronouns, adverbs, conjunctions, and so forth. Each row in $X$ would contain those data for an individual blog post. $\beta$ contains the coefficients for the linear model, and it is a vector of values we solve for when fitting our model to the data. Each item in $\beta$ is a coefficient that defines how each predictor measure is predictive of $y$.

The terms $Z$ and $\gamma$ are the random complements to the fixed effects terms $X$ and $\beta$. Z encodes data about the population or populations that each blog post may belong to, which in the context of this study may include which blog a given post came from or the gender of the blog's author. $\gamma$, like $\beta$, is vector of coefficients. $q$ is the total number of blogs, so in our case $q = 100$.

Lastly, $\varepsilon$ simply accounts for the variance not explained by the data present in $X$ and $Z$.

### *3.4.3 Tools*

For all of our analyses, we used SPSS, Microsoft Excel, and the R statistics package. Microsoft Excel is well-known spreadsheet software that is capable of performing limited statistical analyses but has fairly robust functionality for generating charts and tables. R is a free, open-source, expandable statistics programming language and software environment that has increased in popularity among data miners in recent years. It is operated from a command line, and a number of textual analysis packages for R are available. SPSS is a popular software application for social scientists published by IBM, used for statistical analysis. It handled the large amounts of numerical data much more effectively than Excel.

# Chapter 4 : Demographics and the Language of Political Orientation

## 4.1 Introduction

A growing body of research suggests that differences exist between liberal and conservative Americans that extend beyond their political viewpoints. Academic and journalistic inquiry has in recent decades put for the notion that liberals and conservatives rely on different moral foundations (Haidt and Graham 2007, Graham et. al. 2009), have tendencies towards a different set of personality traits (Hirsh et. al. 2010), watch different television programs (Mitchell et. al. 2014), have different spending habits (Furnham 1985), different brand preferences (Nunberg 2007) and even prefer different beer (Khan et. al. 2013). Amidst widespread reports of deepening political polarization in the United States, attention to differences (and the nature of these differences) between liberals and conservatives has enjoyed renewed attention in the academic world.

A number of such studies have focused on the writings of liberals and conservatives, and the explosion of the popularity of social media in recent decades (for example through blogs) has enabled a line of inquiry on political writings in blogs (Perlmutter 2008, Adamic and Glance 2005, Wallsten 2007, Tremayne 2012, Coleman and Wright 2008, Koop and Jansen 2009). This significant body of study on political blogs has to date focused on the role of blogs in the media ecosystem, the semantic content of the blog itself (for example, the topics discussed, the sentiment of the authors, the opinions the blogs reflect), or on blog metadata (when the blog was published, aspects of the identity of blog authors, and so forth). This chapter of the dissertation will focus on an investigation into patterns in the structure of language used by liberal and conservative bloggers, which have not yet been examined in an academic context. As we describe in prior chapters, recent discoveries reveal correlations between use of function words and different types of observable behavior. Function words are the unobtrusive categories of words such as pronouns, prepositions, and articles, which make up the syntactic structure through which we communicate ideas verbally (Chung and Pennebaker 2007, Pennebaker 2011).

We seek to study these issues through analysis of the dataset identified in Chapter 3, examining function word use in top political blogs collected from May 2012 through December 2012. Our approach bridges previously observed differences in the use of language among various demographics with differences in political tendencies among those demographics. We draw on theory from cognitive science to hypothesize that differences will exist in the way liberals and conservatives fundamentally approach the use of language, and test this hypothesis using an examination of quantitative structural markers of our political blog corpus.

## 4.2 Ideology and Language

It is nearly tautological to say that liberals and conservatives display different voting behaviors, as political parties are often formed around ideologies, with Democrats and Republicans in the United States being no exception (McCarty et. al. 2006, McClosky

1964). Consistent with their voting behaviors, Republicans and Democrats display different and relatively consistent policy preferences (Page and Jones 1979). In recent years, some have suggested that liberal and conservatives might have fundamentally different thought processes, necessarily basing their worldviews on differing sets of "moral foundations" (Shapiro 2012, Graham et. al. 2009). Perhaps unfortunately, these differences are thought to pose significant challenges to mutual understanding and even basic communication between liberals and conservatives.

Why might ideology play a role in how people write? Ideology has long been regarded in many disciplines as the product of logical processes, with ideological stances taken in utilitarian conformity with one's optimal self-interest. Much of modern economics and some veins of modern political science are based on *rational choice* theory, which is largely predicated on this assumption (Black et. al. 1987).

However, critiques of this model have spurred new research on ideology, raising interesting questions regarding the etiology of ideology and what the social and psychological implications of a different understanding of ideology might be. Increasingly, researchers have put credence in the idea of bounded rationality and that deep, underlying psychological factors might influence the ideological orientation of a given individual (Jost et. al. 2009). For example, one study of twins estimated that 40-50 percent of variability in ideological opinions was attributable to genetic factors (Alford et. al. 2005), and another longitudinal study found that the characteristics of childhood personalities had an impact on political orientation later in life (Block and Block 2006).

These findings and others in their vein call into question the cardinality of the rational choice model and suggest that further study of what psychological processes or features might underlie political ideology is needed.

One study (Carney et. al. 2008) offers a taxonomy of various personality types, suggesting that each personality type leads to a different style of verbal interaction. Expressive, excited, enthusiastic, sensitive, and tolerant interaction types are theorized to be stronger among liberals, while interactions characterized as stern, cold, mechanical, withdrawn, reserved, stubborn, or restrained tended to be correlated with conservatives. The study coded both verbal and nonverbal behavior.

Lakoff approaches issues of politics and language with a qualitative textual approach, modeling the country as a “national family,” likening government to parentage and proposing that liberals and conservatives fulfill different roles (Lakoff 2010). According to Lakoff, each role, and thus each political orientation, embodies fundamentally different cognitive models—opposing constructions of morality with fundamentally differing psychological underpinnings. As such, they naturally couch their ideological perspectives in fundamentally different terms with fundamentally different motivations. On the one hand, conservatives embody the “strict father,” who assumes that “life is difficult and the world is dangerous,” while liberals embody the “nurturant parent,” which is centered around mutual respect, caring for others and being cared for. Both moral foundations, he argues, are idealizations that have pervaded the American psyche, and that Americans will be innately familiar with both.

According to Lakoff, the Strict Father model is based on the view that “life is difficult and the world is fundamentally dangerous.” The model embodies the primal experience of a family in which “a father takes primary responsibility for the protection and support of the family,” and “teaches his children right from wrong by setting strict rules for their behavior and enforcing them through punishment… only if a child learns self-discipline can he become self-reliant later in life.” The model presupposes the idea that life is a struggle for survival, and that the moral imperative is success, which is presupposed by self-discipline, obedience to authority, and equated with survival. According to this framework, rewards for those who have not earned them through competition violate the principles of self-reliance and survival, as they “remove the incentive to become self-disciplined and remove the need for obedience to authority.” This model represents the conservative ideology.

Lakoff’s Nurturant Parent Model is an embodiment of the primal experience of “being cared for and cared about, having one’s desires for loving interactions met, living as happily as possible, and deriving meaning from mutual interaction and care.” The nurturant parent seeks to protect their child from the “evils of the world,” whether those be drugs, crime, or pesticides in food or lead paint. This model presupposes empathy—the desire to care for a helpless child implies caring *about* the child. The provision of nurturance is the core moral action, and under this metaphor the citizens of a state are the children, with citizens in need being children in need of nurturance, which may often require sacrifice on the part of the parent. This model is, Lakoff argues, inextricably tied to liberal ideology.

From this framework of moral principles, the presentation of liberal and conservative stances on political issues naturally follows. This is illustrated by approaching how liberals and conservatives describe their policy stances on issues such as gun control, taxes, social programs, and the environment from the perspective of his metaphorical models. For example, under the strict father model, the environment is a gift from God for the stewardship of Man, a source of opportunity and danger, something to be controlled and exploited for the benefit of himself and his children. Under the nurturant parent model, the environment is a provider, a source of sustenance and nurturance in the form of resources, and a victim of predation and abuse by its human caretakers. While a particular political stance is not necessarily determined by the moral inclinations of the policymaker, the manner in which one's political stance is expressed is in terms of these configurations of values and morals. Conservatives and liberals might both express support for environmental protection legislation, for example, but the conservative might present the issue by framing it as promoting sensible and frugal use of available resources, whereas liberals might use more preservationist language that focuses on the intrinsic value of nature and the environment itself. However, though in this particular example the desired outcome is the same, Lakoff points out that more often than not, liberals and conservatives disagree on how the environment should be handled, as it is fundamentally an issue that is beyond "people versus owls or market forces versus the EPA, but [is about] two utterly opposed moral visions of the proper relation of man to nature." This claim and the other parallel claims made about other policy stances have stunning implications—that the mental model individuals have of the proper order of society is subject to a fundamental, dichotomous difference across dimensions of gender and politics.

In general, researchers agree—there is absolutely a connection between ideology and language use. While the mediating psychological processes have been investigated using multiple approaches, we are only beginning to formulate a picture of the nature of that connection.

## 4.2 Demographics and political ideology

It is well known that voting patterns in the United States vary very differently across different sets of demographics. Indeed, demography is frequently considered in political strategy and issues of demography often feature prominently in narratives of elections. We often hear that the results of elections are credited to "whites," "minorities," "young people," "retirees," "women," "the working class," "middle-class voters," and myriad other groups. Some political scientists have alleged that demography, in particular racial demography, plays a primary role in the drawing of congressional district boundaries, with the concept of "racial gerrymandering" having been identified as a sometimes-intentional attribute of representational political processes (Lublin 1999, Forest 2005).

With the idea of demographic differences in behavior taken as a given, some attention has been devoted to investigating the etiology of these observed differences. We review the literature for explanations of differences in voting behavior for three demographic categories: gender, age, and social class.

### *4.2.1 Gender*

Gender and politics are deeply intertwined. Researchers in many fields have found numerous differences in the way that men and women approach politics, in terms of psychology, preferences, and outcomes. A number of papers have found evidence that supports the idea that men and women may conceptualize politics differently. Some attempt to explain voting differences by disentangling various liberal and conservative political stances (Jelen et. al. 1994).

Lakoff insists that both classifications—particularly the "nurturant parent," are not implicitly gendered. He cites the ability of many men to fulfill the nurturant parent role and the ability of women to fulfill the strict father role as a primary reason he avoids characterizing these roles in gendered terms.  Others, however, (Hayden 2003) have argued that the gendered characterizations of "strict father" and "nurturant mother" are in fact the idealizations in the American imaginary to which Lakoff refers, and that while not all who subscribe to these moral frameworks conform to the gender to which they correspond, both sets of moralities are inextricable from the gender role to which they refer. We agree with this characterization. We find that Lakoff's line of argumentation reflects a gender-essentialist worldview, conflating gender-as-identity and gender-as-social-structure. While Lakoff rightly points out that adherents to the “strict father” and “nurturant parent” archetypes can be found in both men and women, and that political sympathies are not cleanly divided by gender lines, this does not preclude the political ideologies themselves from being implicitly gendered or different from the gender of that ideology’s individual adherents. As an example, historical references to various states as “Motherland” or “Fatherland” are certainly at least superficially gendered. Less superficially, the nature of these parental metaphors and of the social construction of gender means that we must consider the “strict father” model masculine and the “nurturant parent” model feminine. While exceptions naturally exist, the definitions of these models and the definition of “conventional” masculinity and femininity have the same origin.

Of the three demographic categories we will examine, research has produced the clearest connections between gender and underlying political psychology. This is particularly fitting because the concepts, in part, overlap: gender itself is socially constructed in a way that wealth or age are not—it, like ideology, is a normative collection of ideas that determines and constrains the way in which society is structured. That gender and political ideology would be closely related or that political ideologies might be inherently gendered seems quite plausible.

### *4.2.2 Age*

It is widely known that voters of different ages groups have significantly different voting tendencies at the polls. One study on immigrant participation in Canadian politics suggests that older individuals are "likely to be stronger political partisans and more politically involved than their younger counterparts," and that "rather than inhibiting participation, age seems to enhance it" (Black et. al. 1987). Advocacy organizations like the AARP popularize awareness of issues faced by older Americans as a group, and have led many to speculate that the promotion of cohort-based interests is a primary motivating factor for older people. In the United States, it is well known that older voters have tended to be more conservative, and it has been speculated that resistance to change characterizes the behavior and ideological tendencies of older voters. While the empirical shape of this result has long been well known in politics, the researchers' study of an immigrant population rejected the theory that a long history of self-reinforcing stimuli motivated older citizens to vote as a matter of inertia. The possibility that resistance to change played a part was also rejected, because despite needing to learn an alien system of politics, older immigrants still participated more than younger immigrants. The study could not explain why, but the authors speculated that family ties and other relational motivations might have been contributing factors.

Other studies have helped to unpack the role of age in explaining the political behavior of older Americans. One investigation that used longitudinal data on voting behavior found evidence that the voting patterns of older people could be to some extent explained by economic and partisan differences (Rhodebeck 1993). A study of ideology in three-generation families found that socialization and status inheritance were found to better account for variance in political beliefs than age (Glass et. al. 1986). The idea that age itself is likely not a significant mediating factor in political beliefs has been supported by at least one study of populations outside the United States (cf Wagner and Kritzinger 2012).

### *4.2.3 Social class*

Social class and ideology have long been inextricable. In addition to being a demographic characteristic, the preferred configurations of social class have invariably been core components of political ideologies. Since the development of the modern concepts of social class, it has featured prominently in narratives of political histories and political theory-crafting. Alignments, realignments, and de-alignments of certain socioeconomic classes with political movements of all stripes have been analyzed and debated at length (cf Brooks and Manza 1997).

In American politics, a narrative exists in which contemporary liberals represent the interests of the poor and marginalized, whereas contemporary conservatives tend to represent those of higher socioeconomic class. Rich voters have been observed in recent years to be far more conservative than poor voters, but rich states have enacted many more liberal, social welfare-oriented policies than poor states (Gelman 2009). To this day, past research has shown “across countries and decades that there exist stable correlations between wealth status and political orientations (Rindermann et. al. 2012). This may be increasingly relevant in the United States, in particular: a Pew Research report in 2012 suggests that a rising share of Americans see a conflict between the rich and the poor (Morin 2012).

Wealth and the redistribution of wealth are certainly one of the primary objects of political ideology, and the implementation of ideologies regarding wealth has readily apparent direct impacts on a populace, with some that quantifiably benefit and some that quantifiably suffer. It is unsurprising that the presence or absence of wealth might produce an element of self-interest in voting. However, it is far from the case that the wealthy always vote in the interests of keeping their wealth—many researchers have found that there is significant variation in political affiliation even among the wealthy. In addition, even from an ideological perspective, a majority of people who self-identified as conservative and wealthy favored, on the whole, a more equal distribution of wealth from the status quo (Norton and Ariely 2011).

A key question remains, though: whether wealth, as some allege, affects both how one votes and how one thinks. Wealthier people, on average, have more stable home lives, are happier, more confident, have more friends, are less anxious, and have better health (Furnham and Argyle 1998). The authors point out that this may have significant effects on how one sees and interacts with the world, and that some underlying psychological processes for the wealthy and the poor may differ as a result. They admit, though, that the connection between money and political orientation is vague at best. Whether and how money itself is a psychological driver of political choice or if there exists some coherent combination of mediating effects remains unknown.

## 4.3 Demographics and Language Use

Pennebaker, in *The Secret Life of Pronouns* (2011)*,* presents extensive studies on large email and blog textual corpora, finding a number of sets of interesting correlations between demographic and social factors and a person's use of language. Using a tool named Linguistic Inquiry and Word Count (LIWC) (Pennebaker et. al. 2007), he found that the frequency of use of various categories of "non-content words"—words like pronouns and prepositions that are the scaffolding through which meaning is conveyed—varied significantly across populations. This work is particularly of note because to date, most textual analysis research had focused on content, while this examined the *structure* of language—an additional dimension of clear import. Variation between other populations were also found, including people of different ages and power.

| **Women use more:** | **Men use more:** |
|---|---|
| Personal pronouns (I, you) | Prepositions |
| Verbs (eat, jump, decide) | Numbers (1, one, twenty-three) |
| Certainty words (absolutely, always) | More words per sentence |
| Hedge phrases (maybe, perhaps, guess) | Big words (longer than six letters) |
| Negations | Articles |
| I-words | |

| **Older people use more:** | **Younger people use more:** |
|---|---|
| Cognitive Words | Personal pronouns |
| Articles | Verbs |
| Prepositions | Time references |
| Positive emotions | Negative Emotions |
| Future-tense verbs | Past-tense verbs |
| Big words | |

| **High social class uses more:** | **Low social class uses more:** |
|---|---|
| Big words | Personal pronouns |
| Articles | First-person singular pronouns |
| Prepositions | Impersonal pronouns |
| | Auxiliary Verbs |
| | Present-tense verbs |
| | Cognitive mechanism words |

**Table 4.1. Correlations between age, gender, social class, and LIWC's linguistic markers (Pennebaker 2011)**

Table 4.1 shows the correlations Pennebaker found between age, social class, gender, and the use of different categories of functional words. Pennebaker notes that statistically, each of these groupings have very large effects. He notes the similarity between each of these groupings, with two clusters of words in particular emerging. The first, the *noun cluster*, includes articles, nouns, prepositions, and big words. The second, called the *pronoun-verb cluster*, includes personal and impersonal pronouns, verbs, auxiliary verbs, and cognitive words.  He notes that men, elderly people, and the rich tend to use more words in the noun cluster, while women, young people, and the poor tend to use more words in the pronoun-verb cluster.

The explanations offered for this remain in the realm of speculation, and as the author admits, “most simple explanations for these differences fall apart at some point.” A number of possibilities are offered—different groups may have a greater or lesser propensity towards certain kinds of thinking—narrative, formal, or analytic. There may be differential power and status between each of the demographic categories, which may have effects in speech. Why and how verbs, for example, may be associated with lower levels of power is unclear.

## 4.4 Hypothesis

The emerging body of research in moral foundations has suggested that underlying psychological mechanisms may exist that explain both the opposing political stances taken by liberals and conservatives in addition to the way in which these mechanisms are verbally justified. We recall that evidence suggests personality traits and moral values are a causative factor in political orientation (Graham et. al. 2009, Block and Block 2006). Qualitative analysis by Lakoff suggests that the way liberals and conservatives present their worldviews is linguistically fundamentally different, and that this presentation of political positions is universally couched in terms of deeply held beliefs that are central to personality and who we are as people.

Meanwhile, discoveries about the use of functional words reveal that different demographics use these words at statistically different frequencies (Chung and Pennebaker 2007, Pennebaker 2011). Findings have supported the idea that psychological features like personality, socialization, and perceptions of power and status may be driving factors in the use of these words. As with political ideology, understanding the underlying psychological mechanisms for differing use of function words remains speculative.

We hypothesize the following:

H1: Use of linguistic markers by liberal bloggers will associate with the specific patterns of linguistic marker use identified in women as detailed by Pennebaker, while use of linguistic markers by conservative bloggers will associate with language used by men.

H2: Use of linguistic markers by liberal bloggers will associate with the specific patterns of linguistic marker use identified in younger people as detailed by Pennebaker, while use of language by conservative bloggers will associate with language used by older people.

H3: Use of linguistic markers by liberal bloggers will associate with the specific patterns of linguistic marker use identified in lower-class people as detailed above, while use of language by conservative bloggers will associate with language used by upper-class people.

## 4.5 Methods

I used the dataset taken from 50 conservative and 50 liberal blogs, obtained as described in Chapter 3 of this dissertation. The posts were each authored between May 1, 2012 and December 31, 2012, bracketing the United States Presidential Election. Over 80,000 unique posts were collected and analyzed.

The blog data we collected serves as a good source of data for the investigation of this topic. The vast majority of posts of the political blogs we identified concerned a single area of attention—politics. The blogs were largely focused on the same subjects because we analyzed posts authored during the election season.

Numerous polls and studies have revealed that in the United States, people who identify as conservatives are likely to be older (Dychtwald 1999), wealthier (Page et. al. 2013), and more likely to be a man (Shapiro and Mahajan 1986). While we were unable to determine the race or age of the authors of the blogs we studied, we attempted to code blog authors by gender. We examined "about the author" pages and blog posts containing personal narratives, and as many blogs had names, pictures, and descriptions that used gender-specific pronouns, we were able to identify the gender of slightly more than half of the blog authors. This exercise revealed that both liberal and conservative bloggers, or at least those identified, were both overwhelmingly men—about eight out of ten identified bloggers were men for both the liberal and conservative blogging populations. Corroborating this result, previous inquiries into political blogging demographics have found remarkable homogeneity in the political blogging population even across political ideologies. A number of studies have confirmed that bloggers, by and large, are overwhelmingly white, well-educated, and male, and that this is relatively invariant across political persuasions (McKenna and Pole 2007, Harp and Tremayne 2006).

To analyze these blogs, I use the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker et. al. 2001, 2007), which is a powerful textual analysis tool that incorporates measures of polarity but also incorporates various affective measures such as sadness or happiness and provides pronoun counts, which have been demonstrated to be correlated to a number of different psychological factors such as perceptions of power, in-group/out-group dynamics, values, and personality (Pennebaker 2011). Such programs have been employed to gain insight into presidential speeches (Bligh et. al. 2004), campaign rhetoric (Olson et. al. 2012), and public opinion (Tumasjan et. al. 2010).

The analysis performed used generalized linear mixed models (GLMM). As previously described, GLMM is an extension of the generalized linear model which takes into account random effects as a part of the linear predictor, whereas generalized linear models only take into account fixed effects. By using GLMM, we are not assuming that any individual specific effect is correlated with the independent variable—which in our case is the binary variable of whether a post is liberal or conservative. Naturally, an approach that is dependent on quantitative computational methods for data analysis has many limitations, among which are included a difficulty in demonstrating causative relationships between metrics for which there is correlation, and the possibility of confounding factors that may remain unknown.

In using GLMM, which we describe at some length in Chapter 3, we defined $y$, the target variable, as the binary of whether an individual blog post was liberal or conservative. Each row of the matrix $X$ corresponds to a single blog post, with each column corresponding to a predictive term, which in our case was the LIWC output for the prevalence of use of certain categories of words, the specifics of which are laid out in the tables below. It is important to note that where LIWC measures are dependent on word counts, the output is in the form of percentages (i.e. the percentage of words in a document that are pronouns). For the benefit of the reader, some descriptive statistics are detailed ahead, in Table 4.5.

As mentioned, each blog was coded as either conservative or liberal. Moderate, bipartisan, or nonpartisan blogs were not collected for the purposes of the study. We coded all posts from liberal blogs as being liberal (l), and all posts from conservative blogs as being conservative (c), not reading each post to make sure it had liberal or conservative content. We identified the blog from which individual posts came in the statistical model (this served as the $Z$ term, with one column each for each blog, and one row each for each blog post, with $Z_{postid,blogid} = 0$ if the post did not belong to the blog and a $Z_{postid,blogid} = 1$) if it did. This was used to control for random effects, eliminating the possibility that one particularly prolific blog with a certain pattern of language use would skew the results.

## 4.6 Analysis

In the tables that follow, Lib/Con refers to whether the measure was found by the model to correlate positively with liberal posts or conservative posts. The model term column contains the list of *p* predictors, which in our case are the characteristics of the blog posts that were measured by LIWC. The "Associated with" column indicates the demographic correlation identified by Pennebaker (2011) of that model term in this specific configuration. The coefficient column is essentially *β,* a list of length *p* that specifies how each model term is predictive of the target variable. Significance is an indicator of whether the model term has a statistically significant effect on the predictiveness of the linear model as a whole.

### *4.6.1 Age*

We tested across the measures indicated by Pennebaker as being associated with old and young people, with the results shown below in Table 4.2. As the reference value for the binomial target variable (liberal vs. conservative) was liberal, positive coefficients meant the term was used more often by conservatives, and negative coefficients meant the term was more used more often by liberals.

| **Lib/ con** | **Model Term** | **Assoc. w/** | **Coeff.** | **Std. Er** | **t** | **Sig.** | **Lower** | **Upper** |
|---|---|---|---|---|---|---|---|---|
| lib | Personal Pronouns | Young | -0.053 | 0.003 | -18.417 | .000 | -0.058 | -0.047 |
| lib | Verbs | Young | -0.045 | 0.002 | -18.217 | .000 | -0.05 | -0.04 |
| con | Negative Emotions | Young | 0.038 | 0.003 | 11.035 | .000 | 0.032 | 0.045 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| con | Time References | Young | 0.016 | 0.002 | 6.721 | .000 | 0.011 | 0.021 |
| con | Past-tense verbs | Young | 0.029 | 0.004 | 7.033 | .000 | 0.021 | 0.038 |
| con | Future-tense verbs | Old | 0.092 | 0.008 | 11.868 | .000 | 0.076 | 0.107 |
| con | Large words | Old | 0.015 | 0.001 | 26.574 | .000 | 0.014 | 0.017 |
| con | Prepositions | Old | 0.037 | 0.002 | 20.346 | .000 | 0.033 | 0.04 |
| con | Cognitive Words | Old | 0.004 | 0.001 | 2.583 | .010 | 0.001 | 0.006 |
| - | Articles | Old | 0.001 | 0.002 | 0.548 | .584 | -0.003 | 0.006 |
| lib | Positive Emotions | Old | -0.008 | 0.003 | -2.999 | .003 | -0.013 | -0.003 |

**Table 4.2. Results of model containing predictor variables for age. Measures colored red were used more frequently by conservatives, and blue more frequently by liberals.**

Of the five measures associated with young people, 2 were used more often by liberals and 3 were used more by conservatives, while of the six measures associated with old people, 1 was associated with liberals, 4 were associated with conservatives, and one had no significant correlation.

### *4.6.2 Social Class*

| **Lib/ con** | **Model Term** | **Assoc. w/** | **Coeff.** | **Std. Er** | **t** | **Sig.** | **Lower** | **Upper** |
|---|---|---|---|---|---|---|---|---|
| con | Large words | Hi Class | 0.016 | 0.001 | 28.005 | .000 | 0.015 | 0.017 |
| con | Articles | Hi Class | 0.004 | 0.002 | 1.705 | .088 | -0.001 | 0.009 |
| con | Prepositions | Hi Class | 0.040 | 0.002 | 22.264 | .000 | 0.037 | 0.044 |
| lib | Personal pronouns | Lo Class | -0.040 | 0.002 | -18.417 | .000 | -0.058 | -0.047 |
| lib | 1st person pronouns | Lo Class | -0.047 | 0.006 | -7.615 | .000 | -0.059 | -0.035 |
| lib | Impersonal pronouns | Lo Class | -0.084 | 0.004 | -23.575 | .000 | -0.091 | -0.077 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| con | Auxiliary verbs | Lo Class | 0.021 | 0.003 | 6.247 | .000 | 0.014 | 0.027 |
| lib | Present-tense verbs | Lo Class | -0.032 | 0.003 | -10.226 | .000 | -0.038 | -0.026 |
| con | Cognitive mechanism | Lo Class | 0.010 | 0.001 | 6.706 | .000 | 0.007 | 0.013 |

**Table 4.3**. Results of model containing predictor variables for social class. **Measures colored red were used more frequently by conservatives, and blue more frequently by liberals.**

We tested across the measures Pennebaker identified as being associated with people of lower socioeconomic class and those of higher socioeconomic class. As the reference value for the binomial target variable (liberal vs. conservative) was liberal, positive coefficients meant the term was used more often by conservatives, and negative coefficients meant the term was more used more often by liberals. Of the three measures associated with higher socioeconomic class, conservatives used significantly more large words and prepositions. Use of articles showed a trend (sig. = 0.88) towards more frequent use by conservative. Of the six measures associated with individuals with lower socioeconomic class, four were associated with liberals and two with conservatives. All correlations were found at a significance of less than 0.05, except personal pronouns, which showed a trend at a significance of 0.088.

### *4.6.3 Gender*

We tested the factors associated with gender below using GLMM. In the table below, blue indicates that liberals had a greater usage of the measure, and red indicates that conservatives had more. For the most part, the pattern of language use across genders identified by Pennebaker is matched, with the exception of articles and negations—there was no significant correlation for those two categories. All other correlations matched the L-feminine C-masculine breakdown identified by Pennebaker.

| lib/ con | Model Term | Assoc. w/ | Coeff. | Std. Error | t | Sig. | Lower | Upper |
|---|---|---|---|---|---|---|---|---|
| con | Prepositions | Men | 0.034 | 0.002 | 17.590 | .000 | 0.030 | 0.038 |
| con | Numbers | Men | 0.019 | 0.007 | 2.886 | .004 | 0.006 | 0.032 |
| con | Words per sentence | Men | 0.006 | 0.000 | 11.889 | .000 | 0.005 | 0.006 |
| con | Large words | Men | 0.016 | 0.001 | 25.811 | .000 | 0.015 | 0.017 |
| - | Articles | Men | .0002 | .0003 | 0.774 | 0.439 | -0.029 | 0.007 |
| lib | Personal pronouns | Women | -0.038 | 0.003 | -10.869 | .000 | -0.045 | -0.031 |
| lib | Verbs | Women | -0.025 | 0.002 | -11.146 | .000 | -0.030 | -0.021 |
| lib | Certainty | Women | -0.016 | 0.006 | -2.671 | .008 | -0.029 | -0.004 |
| lib | Hedge words | Women | -0.015 | 0.005 | -3.013 | .003 | -0.024 | -0.005 |
| - | Negations | Women | -0.047 | 0.006 | 0.610 | .542 | -0.008 | 0.016 |
| lib | First-person pronouns | Women | -0.047 | 0.007 | -7.245 | .000 | -0.060 | -0.035 |

**Table 4.4. Results of model containing predictor variables for gender. Measures colored red were used more frequently by conservatives, and blue more frequently by liberals.**

More specifically, as the reference value for the binomial target variable (liberal vs. conservative) was liberal, positive coefficients meant the term was used more often by conservatives, and negative coefficients meant the term was more used more often by liberals. Of the 5 measures associated with men, 4 were used more often by conservatives and one had no significant correlation. Of the 6 measures associated with women, 5 were used more often by liberals and one had no significant correlation. Of the 9 measures with significant differences in usage between the two populations, the correlation coefficient was lower than 0.01, and .001 or lower for six of the measures.

| Liberal / conservative | | Min | Max | Mean | Std. Deviation |
|---|---|---|---|---|---|
| Con. | Prepositions | .00 | 50.00 | 13.2068 | 4.38630 |
| | Number words | .00 | 40.00 | .8863 | 1.26456 |
| | Words per sentence | 1.00 | 1752.00 | 25.2673 | 23.28006 |
| | Large words | .00 | 100.00 | 28.0868 | 19.68939 |
| | Articles | .00 | 40.00 | 7.2848 | 3.40838 |
| | Personal pronouns | .00 | 50.00 | 3.6861 | 2.97633 |
| | Verbs | .00 | 66.67 | 9.1949 | 4.31568 |
| | Certainty words | .00 | 33.33 | 1.0812 | 1.37675 |
| | Hedge words | .00 | 100.00 | 1.7947 | 1.88734 |
| | Negations | .00 | 100.00 | 1.1047 | 1.44656 |
| | 1st person pronouns | .00 | 33.33 | .7123 | 1.42239 |
| Lib. | Prepositions | .00 | 60.00 | 12.2337 | 5.18673 |
| | Number words | .00 | 50.00 | .8169 | 1.15508 |
| | Words per sentence | 1.00 | 1723.00 | 22.1004 | 22.33565 |
| | Large words | .00 | 100.00 | 25.1551 | 8.85045 |
| | Articles | .00 | 37.50 | 6.9024 | 3.69312 |
| | Personal pronouns | .00 | 66.67 | 4.1448 | 3.48918 |
| | Verbs | .00 | 75.00 | 9.5918 | 5.15630 |
| | Certainty words | .00 | 50.00 | 1.1237 | 1.35660 |
| | Hedge words | .00 | 50.00 | 1.8562 | 1.78377 |
| | Negations | .00 | 40.00 | 1.1568 | 1.39969 |
| | 1st person pronouns | .00 | 50.00 | .9394 | 1.82128 |

**Table 4.5. Descriptive statistics for predictor variables for gender**

Of the three models, this one most clearly shows that the prevalence for liberal and conservative use of these categories of function words most clearly matches that found across male and female use of those same words.

SPSS does not provide an $r^2$ value for GLMM, but instead characterizes the model generated as having some degree of accuracy in predicting the target variable (in our case, whether a blog post is liberal or conservative). This characterization revealed that our model is 55.8% accurate. Descriptive statistics for the predictor variables for gender are shown above in Table 4.5.

With the reference category of l, a negative coefficient indicates that a given factor (labeled as "model term") was more likely to appear in liberal blogs, whereas a positive number would indicate that it was more likely to appear in conservative blogs.

### *4.6.4 Controlling for Author Gender*

Naturally, one must consider whether our result, which involves the correspondence of patterns of linguistic production among liberals and conservatives with those patterns previously observed among men and women might be mediated by author gender. The proportion of men and women might differ between liberal and conservative blogs, and this difference in proportions might be enough to account for the linguistic patterns observed above. In such a case, rather than reinforcing Lakoff's connection between gendered psychology and ideological stance, our findings would serve only to reaffirm the findings that men and women use function words in different ways. While a 2007 survey of political blogs (McKenna and Pole) found that 75% of blog authors were men and that this was consistent across liberal and conservative blogs, with the small effect sizes we observe, it is possible that small differences in blogger gender could account for the variance in language usage.

For many of the blogs, identifying author gender was simple as they were written by a single author, whose gender identity was publicly displayed online. However, for other blogs this presented difficulties, as several had multiple contributors who authored posts during our data collection period. In addition, some bloggers operated under pseudonyms, making it difficult or impossible to identify their gender. To address this, we coded blog posts for authorship by men or women. To do so, we employed the following protocol:

1) Determine whether the blog has an individual or multiple authors.
2) If the blog has a single author, determine and code gender of that author by:
    a. Examining use of self-referential gendered pronouns.
    b. Finding other explicit references to author gender, either in blog posts by the author or in the "About" section of the website.
    c. Determining whether the name of the author is gendered in standard American usage.
    d. If author gender is still indeterminate, code that author as having an "Unknown" gender. Many bloggers used pseudonyms and did not otherwise identify their gender.

3) If the blog has multiple authors, identify all authors that posted during our data collection timeframe. Use steps 2a through 2d above to code the gender of each author.

This process produced the following results, in aggregate:

| | #Men | %Men | #Women | %Women | #Unknown | %Unknown |
|---|---|---|---|---|---|---|
| Liberal | 196 | 65.1% | 61 | 20.3% | 44 | 14.6% |
| Conservative | 170 | 73.0% | 42 | 18.0% | 21 | 9.0% |

**Table 4.6. Gender Counts of Blog Authors**

We then created a scale, which we called *masculinity.* Each blog post was assigned a masculinity rating based on the gender identity of its authors. For blogs with single authors who identified as men, the posts belonging to those blogs were given a masculinity value of 1. For blogs with single authors who identified as women, we coded posts belonging to those blogs with a masculinity value of 0. For blogs with multiple authors, rather than attempting the impractical task of labeling each blog post with the gender of the individual author, we defined the blog's masculinity as the proportion of contributors to a blog who were men.

We then employed three strategies to deal with the bloggers with unknown gender. The first method, which we will call "masculinity-75" we used assumed that the demographics of the unknown bloggers matched those reported of political bloggers generally by McKenna and Pole (2008)—seventy-five percent male. The second method, which we will call "masculinity-50" assumes that the gender composition of the bloggers operating with a pseudonym is evenly distributed between men and women. The final method, "masculinity with unknowns excluded" omitted the bloggers we were unable to code for gender entirely from our calculation of the masculinity measure. Where all bloggers were of unknown gender, no data for masculinity was recorded when the second method was used. The formulae we used were as follows:

$$masculinity_{75} = \frac{n_{men} + 0.75 * n_{unknown}}{n_{men} + n_{women} + n_{unknown}}$$

$$masculinity_{50} = \frac{n_{men} + 0.5 * n_{unknown}}{n_{men} + n_{women} + n_{unknown}}$$

$$masculinity_{unknowns\_excluded} = \frac{n_{men}}{n_{men} + n_{women}}$$

We then conducted two more GLMM analyses, each incorporating one of the masculinity measures as random effects to see if changes occurred in the significance of fixed effects. If the gender identity of bloggers mediated the result we found, we would expect to see the significance of the fixed effects reduce. We obtained the following results:

| **lib/ con** | **Model Term** | **Assoc. w/** | **Coeff.** | **Std. Error** | **t** | **Sig.** | **Lower** | **Upper** |
|---|---|---|---|---|---|---|---|---|
| - | Prepositions | Men | 0.005 | 0.003 | 1.815 | .069 | 0.000 | 0.010 |
| - | Numbers | Men | 0.014 | 0.009 | 1.599 | .110 | -0.003 | 0.031 |
| con | Words per sentence | Men | 0.001 | 0.000 | 3.167 | .002 | 0.000 | 0.002 |
| con | Large words | Men | 0.035 | 0.001 | 24.428 | .000 | 0.032 | 0.037 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| con | Articles | Men | 0.038 | 0.003 | 10.990 | .000 | 0.031 | 0.045 |
| lib | Personal pronouns | Women | -0.021 | 0.005 | -4.246 | .000 | -0.030 | -0.011 |
| lib | Verbs | Women | -0.045 | 0.003 | -15.207 | .000 | -0.051 | -0.039 |
| lib | Certainty | Women | -0.041 | 0.008 | -5.278 | .000 | -0.056 | -0.026 |
| lib | Hedge words | Women | -0.076 | 0.007 | -10.963 | .000 | -0.090 | -0.063 |
| lib | Negations | Women | -0.026 | 0.008 | -3.300 | .001 | -0.042 | 0.011 |
| lib | First-person pronouns | Women | -0.027 | 0.010 | -2.798 | .005 | -0.046 | -0.008 |

**Table 4.7. Results of model containing predictor variables for gender, with the masculinity-75 measure incorporated as a random effect.**

| lib/ con | Model Term | Assoc. w/ | Coeff. | Std. Error | t | Sig. | Lower | Upper |
|---|---|---|---|---|---|---|---|---|
| - | Prepositions | Men | 0.003 | 0.003 | 1.063 | .288 | -0.002 | 0.08 |
| - | Numbers | Men | 0.009 | 0.009 | 1.058 | .290 | -0.008 | 0.026 |
| con | Words per sentence | Men | 0.002 | 0.000 | 3.912 | .000 | 0.001 | 0.002 |
| con | Large words | Men | 0.033 | 0.001 | 23.687 | .000 | 0.031 | 0.036 |
| con | Articles | Men | 0.037 | 0.003 | 10.657 | .000 | 0.030 | 0.044 |
| lib | Personal pronouns | Women | -0.024 | 0.005 | -4.944 | .000 | -0.033 | -0.014 |
| lib | Verbs | Women | -0.040 | 0.003 | -13.522 | .000 | -0.046 | -0.035 |
| lib | Certainty | Women | -0.038 | 0.008 | -4.895 | .000 | -0.054 | -0.023 |
| lib | Hedge words | Women | -0.077 | 0.007 | -10.998 | .000 | -0.091 | -0.063 |
| lib | Negations | Women | -0.023 | 0.008 | -2.822 | .005 | -0.038 | 0.007 |
| lib | First-person pronouns | Women | -0.034 | 0.010 | -3.524 | .000 | -0.053 | -0.015 |

**Table 4.8. Results of model containing predictor variables for gender, with the masculinity-50 measure incorporated as a random effect.**

| lib/ con | Model Term | Assoc. w/ | Coeff. | Std. Error | t | Sig. | Lower | Upper |
|---|---|---|---|---|---|---|---|---|
| - | Prepositions | Men | -0.002 | 0.003 | -0.677 | .499 | -0.007 | 0.004 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| - | Numbers | Men | 0.009 | 0.009 | 1.008 | .314 | -0.008 | 0.026 |
| con | Words per sentence | Men | 0.002 | 0.000 | 4.225 | .000 | 0.001 | 0.003 |
| con | Large words | Men | 0.032 | 0.001 | 21.650 | .000 | 0.029 | 0.035 |
| con | Articles | Men | 0.034 | 0.004 | 9.592 | .000 | 0.027 | 0.041 |
| lib | Personal pronouns | Women | -0.037 | 0.005 | -7.394 | .000 | -0.047 | -0.027 |
| lib | Verbs | Women | -0.031 | 0.003 | -9.956 | .000 | -0.037 | -0.025 |
| lib | Certainty | Women | -0.034 | 0.008 | -4.142 | .000 | -0.050 | -0.018 |
| lib | Hedge words | Women | -0.076 | 0.007 | -10.540 | .000 | -0.090 | -0.062 |
| - | Negations | Women | -0.010 | 0.008 | -1.288 | .198 | -0.026 | 0.005 |
| lib | First-person pronouns | Women | -0.040 | 0.010 | -4.046 | .000 | -0.060 | -0.021 |

**Table 4.9. Results of model containing predictor variables for gender, with the masculinity with unknowns excluded measure incorporated as a random effect.**

For the sake of simplicity, where a measure that was previously identified by Pennebaker as being associated with women's speech is identified by our models as being more prevalent within liberal blogs (or the same with men's speech and conservative blogs), we will say it displays *gender-politics correspondence*. Results for all three models which controlled for gender were reasonably consistent with our earlier results without those controls.

With masculinity-75 and masculinity-50 included as random effects in the model, nine out of eleven predictor variables displayed gender-politics correspondence, though which predictors were significantly predictive and which were not was altered when the random effects were incorporated. Negation words, which were not significant without blogger gender as random effects, were significantly associated with liberal speech. On the other hand, prepositions, which were associated with conservatism when blogger gender was not incorporated in the model, did not show a significant effect. Again, those measures that did not display gender-politics correspondence were not matched differently between gender and politics, but were instead below the threshold of significance.

With the “masculinity with unknowns removed” measure as a random effect, eight out of eleven predictor variables displayed gender-politics correspondence. Prepositions, number words, and negations all did not display significant effects.

## 4.7 Discussion

The data reveal little support for our hypothesis *H1*, as the measures associated with the young and old were not each also clearly associated with liberals and conservatives. *H2* had slightly better support, with 2 of 3 measures associated with high class also correlated with conservatism, and 4 of 6 measures associated with low class also correlated with liberalism. However, *H3* had the strongest support, with 4 of 5 measures associated with men also being correlated with conservatism, and 5 of 6 measures associated with women being correlated with liberalism. The measures that did not display such a gender-politics correspondence were thus not used more often by either liberals or conservatives, at a statistically significant level.

We confirm that significant partisan differences can be discerned in the use of function words in political speech. While language, gender and politics are clearly inextricable, it was not a foregone conclusion that liberals and conservatives would have significant differences in the way they approach speech. There is much variability in the use of function words, with the large body of research we previously referenced having found that factors such as truthtelling, emotion, sex, age, power, and individual style all affect how much these words get used. Political orientation could have been one of the many unmentioned factors that do not affect this unconscious aspect of language. However, our finding shows strong statistical evidence that it *does*—what remains to be done is a grounded speculation as to *why*.

It is important to note that there was some repetition in the measures tested in each of the three models, with personal pronouns, for example, having some salience to demographic differences in language use. This was also noted in *The Secret Life of Pronouns* (2011), which posited that perhaps a factor like power likely underlay at least some of the differences in language use he observed. While that hypothesis may have some merit, we want to focus in particular on the result that provides a strong case that we cannot reject *H3*. If we do not reject H3, we are left to conclude that conservatives may write more like men and liberals may write more like women.

We recall our discussion of Lakoff's model of political ideology, in which a gendered model of morality and values determines how people think and speak politically. We propose that in the exposition of one's ideology, as observable in political blogs, the adherence to the dichotomous roles espoused in Lakoff's framework resulted in the observable effects in the language the bloggers produced.

Controlling for the possibility that the gender breakdown of the blogging population was a mediating effect was essential. While the improbability of coding every blog post for the gender of the author limited our ability to control for author gender, our use of multiple methods to control for author gender and the generally consistent results with our earlier findings suggest that it is less likely that the gender of the authors was a mediating factor in the differences in language we were able to observe.

### *4.7.1 Limitations*

Our first key limitation comes with our inability to label every blog post with the gender of its author—this would provide the most accurate control for the potential influence author gender might have had on our results. We needed to assume that on the whole, the women were not significantly more prolific than the men, not producing more or fewer blogs on average than their male counterparts within a given blog.

Again, the complexity of factors that might influence use of function words presents a significant limitation to our ability to make theoretical claims about our results. We wish to highlight the areas in which controls that could be used in future studies. For one, Lakoff's analysis deals with political speech, as does our data set. If data could be gathered that distinguishes between political and non-political speech in addition to the identity of the speaker or writer, we could find whether it is the act of speaking politically that triggers the identified differences in function word use, or if function word use simply reflects how people use words in everyday speech.

Much as the explanations proposed for the differences in political speech on the part of liberals and conservatives, differences in the use of function words have been presented as being related to psychological processes fundamental to who we are. As mentioned previously, it is still unclear exactly why some people use more nouns and articles, and others use more verbs and prepositions. These speech differences have been speculatively characterized as analytical and narrative speech, respectively, but what circumstances might lead one to use analytical speech as opposed to narrative speech is uncertain.

### *4.7.2 Alternative Explanations*

It is plausible that other mechanisms are at work in producing the result we obtained. Rather than reflecting the moral framework of the bloggers themselves, the explanation could be demographic in nature—it could be that bloggers are targeting their audiences by using language that reflects their audience makeup. As a greater proportion of women are liberal in the United States, while a greater proportion of men are conservative, blog authors could be targeting their audiences by mirroring the linguistic characteristics of their speech. Though women did not make up more of the liberal blogging population than the conservative blogging population, it is possible that blog authors targeted female audiences by talking more like women. Evidence exists that this behavior is plausible: it has long been believed that gender-based differences in language use persisted regardless of audience (Lakoff 1972). However, one experimental study found that both men and women spontaneously used language more similar to that of an opposite-gender conversational partner when compared with when they corresponded with a same-gender conversational partner (Thomson et. al. 2001). Though questions remain as to whether a multidirectional relationship exists between gender of speaker and audience, accounts have been consistent over the past several decades that gender and word use are deeply intertwined (Pennebaker et. al. 2003).

However, in this case we find the argument that the differences we observed are demographically driven to be unconvincing. While, as we discuss in the next chapter, it is plausible that bloggers might adjust their language in response to an audience, it seems farfetched to suggest that authors of political blogs are implicitly targeting women or men. Intriguingly, Lawrence et. al. (2010) found in a study of over 16,000 American political blog readers that there is no significant difference in the readership of liberal and conservative blogs across gender, age, income, or education, an account corroborated by at least one other study (Eveland and Dylko 2007).

Moreover, our data itself does not support the idea that differences in audience demographics mediated the results we obtained, as our models produce results inconsistent with this explanation, with liberal and conservative language displaying differences in function word use that correspond with gender but not age or social class. If bloggers were appealing to a certain gender based on the demographics of their audience, would they not also attempt to appeal to an age group in the same way?  While we cannot definitively determine the demographics of the readership of the blogs we studied, both our results and extant demographic studies of a similar set of blogs suggest that it is unlikely that blog authors are modifying their language to cater to a specific audience.

### *4.7.3 Language, Gender, and Politics*

Lakoff suggests that the underlying morality and view of family in the discourse of American liberal and conservative ideologies includes a strong gendered component. His framework posits that the liberal ideology is connected with features that society identifies as feminine—nurturing, caring, forgiving, and can be identified with a "nurturing mother" figure. Conservative ideologies are identified as congruent with a "strict father" metaphor—rules-, hierarchy-, and security-oriented. We find these results especially interesting in light of a lack of significant difference in gender between conservatives and liberals.

Lakoff's model, popularly cited by academics in linguistics and political science, has (as mentioned previously) found some tentative support through quantitative studies. The result we obtained strengthens these findings. The model was originally developed through analysis of the discourse of liberals and conservatives and their policy positions through the lens of culture. By building on Pennebaker's work and using the LIWC tool, we find that the language use of liberals and conservatives matches Lakoff's gender paradigm, an unintuitive result considering the homogeneously male nature of both the liberal and conservative blogging populations.

The connection between language and gender has long been a subject of inquiry, and the role of men and women in societies has had a profound impact on the use of language and every aspect of its structure, whether it be morphological, syntactic, phonological, or lexical. Languages around the world have gendered parts of speech (such as pronouns, nouns, and adjectives) and even gendered alphabets. Even in languages like English, in which grammatical gender has slowly faded over the centuries, gender differences in language use are profound. Robin Lakoff, in her seminal *Language and Woman's Place* (1975), argued for the existence of a "woman's register," a use of language that fundamentally reflected a submissive and subservient role to men. Men, she argued, were able to be assertive in their speech in a way that was denied to women. Many subsequent empirical researchers and theorists have built on her theory, examining the etiology and various impacts of gender and language in interpersonal relationships, psychological development, cultural norms, and the formation of modern institutions (Freeman 1996; Holmes and Meyerhoff 2006).

The performance of gender, as something that is enacted as a part of everyday social practice (West and Zimmerman 1987), naturally includes the use of language (Eckert 2003). A prevailing theoretical framework for gender and language models language as a practice within a community of practice, in which language is the product of a jointly negotiated practice of gender. This perspective does not presuppose differences between genders and attempt to explain all differences in language as a result of some characteristics of gender, but rather accounts for differences in the performance of gender across different groups. The ways in which gender is expressed through language may differ greatly between different populations and contexts, even within the same language.

With this important caveat, the advent of large-scale textual analysis on digital media has yielded many studies on language use and gender differences in various populations. For example, content analysis has been used to examine gender roles, with one study revealing that male teenage bloggers tended to use language in a more “active and resolute” manner, while showing that there was some convergence in the content of male and female adolescent bloggers (Huffaker and Calvert 2006). Other studies, on the contrary, found stark differences in the vocabulary and content across genders, with women using more emotional language and men using more expletives (Schwartz et. al. 2013). Textual analysis has even produced results that seem to corroborate the communities of practice framework of gender performance in language, finding that individuals whose language does not match others of their gender tend to have fewer same-gender social connections (Bammen et. al. 2014).

Though the above studies on gender and language deal with individual gender differences and use of language, Lakoff’s presentation of a dichotomy in political ideologies identifies ways in which the use of language for political purposes is fundamentally gendered. That gender and politics might be linguistically connected has found some empirical support, with one study in particular confirming that the moral framework he proposes may have some merit (McAdams et. al. 2008). Interviews with liberals and conservatives about their personal lives were extensively coded for expressed moral values, each of which corresponded to a liberal or conservative worldview. In these self-narrations, it was found that conservatives “tended to depict authority figures as strict enforcers of moral rules,” while liberals “were more likely to identify lessons learned regarding empathy and openness.” Though the results of this study were not a complete validation of Lakoff’s predictions, it presented a convincing account of his mapping of values to ideology from a quantitative perspective.

If Lakoff is correct, implications exist for our understanding of political orientation—in particular, it highlights the central role that discourse has in both revealing and determining political orientation. The moral and normative arguments used by people to make political statements are based on sets of principles that follow logically from identifiable, gendered metaphors. To rephrase this: political speech is not only different in character when the particular policy positions espoused are different; rather, the same policy position espoused by individuals of different political persuasions would have linguistic features that are necessarily distinct, because the way liberals and conservatives understand the world is fundamentally different.

We propose that the differences we find in function word use between liberals and conservatives in fact reflect the differences in the moral foundations. The Strict Father metaphor is linguistically represented in a manner similar to masculine speech with regards to the use of function words, while the Nurturant Parent metaphor is represented in a manner more similar to feminine speech. The Strict Father metaphor, for example, is described as having a black-and-white binary approach towards morality (Lakoff 2008), whereas the morality of the Nurturant Parent is described as being more context-dependent. A social orientation is clearly indicated in the Nurturant Parent metaphor, with social nurturance one of the core moralities of the liberal worldview. The conservative worldview, on the other hand, is embedded with a power orientation, with the idea of structure and authority as morality.

While we draw parallels between the use of language by men and conservatives, and women and liberals, we note that we must take care not to equate gender with political orientation. First, though as of late the number of women who vote for liberal politicians has increased, it is still the case that plenty of men and women identify with a variety of political ideologies. When we examine the impact of gender in text, we are not necessarily making claims about the gender roles of the writers, but rather, the tendency of certain types of roles to take on certain types of moralities. The biggest assumption we make is that the morality underlying Lakoff's parenthood metaphor in some way mirrors the differences found in the way women and men understand the world. This, however, may not be so farfetched—Men and women have long had moral differences attributed to them (Kohlberg 1981, 1984; Gilligan, 1977, 1982), which theorists have alternately presented as socialized or innate. Indeed, one of the most popular claims regarding gender and moral orientation is, phrased in general terms, that women have more of an orientation towards care and that men have more of an orientation towards justice (Walker 2006). This claim has been explored at length and has been largely confirmed, though the rationale for this continues to be a matter of some dispute. On a similar vein, textual approaches to analyzing gender and morality reveal that women tend to have more of a social orientation and men tend to have more of a power orientation (Dovidio et. al 1988, Newman et. al. 2008).

In summary, in this context it would seem that the puzzling result that conservatives write more like men and liberals write more like women is consistent with a sociological framework of gender differences. In addition to confirming that there are significant partisan differences in the use of function words, our finding lends support to the veracity of Lakoff's account of the Strict Father and Nurturant Parent model of liberal and conservative worldviews. This finding contributes to part of a broader academic narrative that suggests partisanship cannot be adequately explained by the public choice model, and that ethical motivations orthogonal to popular conceptualizations of rationality are the fundamental driving force behind political orientation and political choice.

# Chapter 5 : Linguistic Style and Political Orientation

## 5.1 Background

It is well understood that many differences exist between liberal and conservative Americans that extend beyond their political viewpoints. In Chapter 4, we describe a litany of differences identified by researchers between liberals and conservatives: moral foundations (Haidt and Graham 2007, Graham et. al. 2009), personality traits (Hirsh et. al. 2010), television programs (Mitchell et. al. 2014), different spending habits (Furnham 1985), different brand preferences (Nunberg 2007) and even different beer preferences (Khan et. al. 2011). Amidst widespread reports of deepening political polarization in the United States, attention to differences (and the nature of these differences) between liberals and conservatives has enjoyed renewed attention in the academic world.

A number of such studies have focused on the writings of liberals and conservatives, and the explosion of social media (for example through blogs) has enabled a line of inquiry n political writings in blogs. Through textual analysis of blogs, it has been found that liberals and conservatives use different words to describe similar phenomena (Iyengar 1990, 1993, 1995, Fine 1992), referred to as *framing*. Various kinds of textual analysis have also examined the interrelationship between blogs, including linking (Adamic and Glance 2005) and quoting (Mullen and Malouf 2006) behavior in political blog comments, finding that political orientation has significant effects on such behaviors.

This body of studies of political blogs have to date have used as the subject of analysis the role of blogs in the media ecosystem (cf Perlmutter 2008, Wallsten 2007), semantic content such as the sentiment of the authors (Tumasjan et. al. 2010), the opinions the blogs reflect (Drezner and Farrell 2004), or blog metadata such as author identity (cf Larsson and Moe 2012). This chapter of the dissertation will focus on an investigation of the *structure of language* used by liberal and conservative bloggers, which has not yet been examined in an academic context. We draw on psycholinguistic theory to hypothesize that differences will exist in the way liberals and conservatives fundamentally approach the use of language, and test this hypothesis using an examination of quantitative structural markers of our political blog corpus.

## 5.2 Linguistic Style

Linguistic style has long been considered a tool for social evaluation, both in the study of linguistics and in everyday life. In the Hebrew Bible, the Gileadites, upon inflicting a military defeat on their Ephramite enemies, used the pronunciation of the word *shibboleth* to determine friend from foe and who might live or die. Centuries later, in *My Fair Lady,* Eliza Doolittle finds that her diction and grammar are inextricably tied to issues of class and social mobility in early twentieth century England. It has been well established that the *manner* in which we convey ideas is often as or more important than what we convey.

Today, linguistic style is a sociolinguistic concept encompassing many ideas. The word 'style' invokes a holistic, subjective impression of the characteristics of a given unit of speech, but has been also been more technically defined in a number of ways that can be generally summarized as the phonological, lexical, syntactic, prosodic, and orthographic differences observed within a single language, as utilized by individuals and groups. Our understanding of linguistic style and the contexts that modulate its production remains incomplete and at times inconsistent. Style has been characterized as at once fluid and contextual (Labov 1990, Bell 1984) and deeply ingrained and invariant across time (Niederhoffer and Pennebaker 2002), depending on the context of analysis. Pennebaker, in *The Secret Life of Pronouns* (2011) offers that "People are inclined to match conversational partners in style, regardless of their intentions and reactions," but later reminds us that "people don't talk the same way unless they are joined together in some common purpose, have common goals, lives, desires." This seemingly paradoxical set of attributes deserves further exposition.

Sociolinguist William Labov introduces the idea of style in *The Social Stratification of English in New York City* (1964), which describes an extensive study of English phonology among people belonging to different socioeconomic classes. He found that pronunciations were consistent within social groups but that there were striking differences across class. Critically, he also observed something he called "style-shifting," finding that the pronunciation changed depending on the context of speech, the audience and the formality of the setting. He writes that there is no single-style individual, finding that just as striking differences exist in the way people from different walks of life produce language, that people invariably and deliberately change the sounds they use to speak, and that these changes involve changes in attention as they relate to speech production.

Giles and Powesland expand the scope of examination of style from the phonological to the lexical and syntactical, and take a closer look at style-shifting in *Speech Style and Social Evaluation* (1978). Basing their work on attribution theory, they focus on the influence that auditor and speaker have on one another, proposing a theory they call *linguistic accommodation*. Sometimes referred to as communication accommodation, the theory holds that participants in a conversation tend to unconsciously alter their language to be more similar to a partner's communicative behavior. More formal speech by one partner is met with more formal speech by the other, and accents and dialects sometimes begin to change to match that of the other speaker. It is proposed tentatively that this is a function of cooperation and the human drive to relate, understand, and be understood (Giles and Smith 1979). It is later found that the linguistic characteristics of the speech of interlocutors tend to converge on a number of metrics, including pauses, diction, syntax, pitch, and gesticulations, and that this is negotiated nearly immediately (Giles, Coupland, Coupland 1991).

In an attempt to more thoroughly explain this accommodation behavior, Bell (1984) proposes that style is a matter of speech design for an audience—that any variations in a person's speech are deliberate and targeted at interlocutors, or auditors, be they individuals or groups. Style, he says, changes in response to situational shifts—as the addressee changes in the course of a conversation, style changes. The addition of auditors (people openly listening to a speaker) to a dialogue has effects on style, as does awareness of overhearers (people who have the ability to listen to a speaker, of whom the speaker may or may not be aware). Bell illustrates the exigency of these style shifts by invoking President Carter's interview with *Playboy* magazine, in which he departed from his deliberate, Presidential interview style and used more casual, perhaps scandalous references to "shacking up." His lack of accounting for the "overhearers" endemic to mass-media publications was an example of poor audience design, a social process which we negotiate every day. It is suggested that linguistic style is a crucial tool in the execution of Goffman's (1959) notion of the everyday performance of self. Intriguingly, Bell further acknowledges the usefulness of style to "redefine the existing situation," facilitating the ability of a speaker to modify a social context, for example making formal interactions more casual and friendly, or vice versa. In this way, style is shown to have both reactive and proactive functions in verbal communication.

The cognitive aspects of accommodation have been more closely examined in recent years in the form of *linguistic alignment*, which describes the phenomenon in which many aspects of linguistic representation in a dialogue seem to *automatically* align as a prerequisite to the act of communication (Pickering and Garrod 2004). Under this framework, the semantic and lexical representation we identified above is reflexively interconnected with both a speaker and a receiver's *situation model*—a multi-dimensional representation of the situation under discussion. Situation models are psychological constructs that encode time, space, intentionality, causality, and other contextual information under discussion (Zwaan & Radvansky 1998). They are the subjects of a large body of research which has shown that when various aspects of situation models are manipulated, text comprehension can be affected—for example, when a word is used that is spatially or temporally related to the topic of discussion, people are faster to recognize it. Spoken plainly, situation models are an attempt to capture what people are “thinking about” as they speak, write, read, or listen.

We believe the theory of linguistic alignment to be applicable to the study of blogs. Linguistic alignment holds that the production process of communication both informs and is informed by a situation model, and that the representation of a message—in the case of the blog, through text—is, as such, influenced by one’s worldview. Situation models have a bidirectional relationship with the language generated by a speaker or writer—while it is obvious that what one chooses to write is determined by the context in which they are writing and the way they see the world, studies have shown, for example, that it is possible to prime various aspects of situation models through framing. For example, having a speaker verify a temporal or spatial reference frame (e.g. A: “it’s on the right, yeah?” B: “Yeah”) will often cause that speaker to adopt that reference frame (Warren and Rayner 2004). Many types of semantic priming are possible, not just temporal-spatial, but also conceptual.

The influence the blogger’s situation model has on the text she creates is limited not only to semantic choices—the syntactic, structural aspects of communication can be indirectly influenced as well, as seen in the figure below. Syntax, according to Pickering and

Garrod (2004) and others, is bi-directionally interconnected with the semantic representation (the meaning the speaker wishes to convey), the lexical representation (the semiotic choices made by the speaker), and, in the spoken word, the phonological representation (the encoding of semiotics and syntax in sound). In the case of the written word, syntactic and lexical representations can be understood to be encoded in a textual representation.

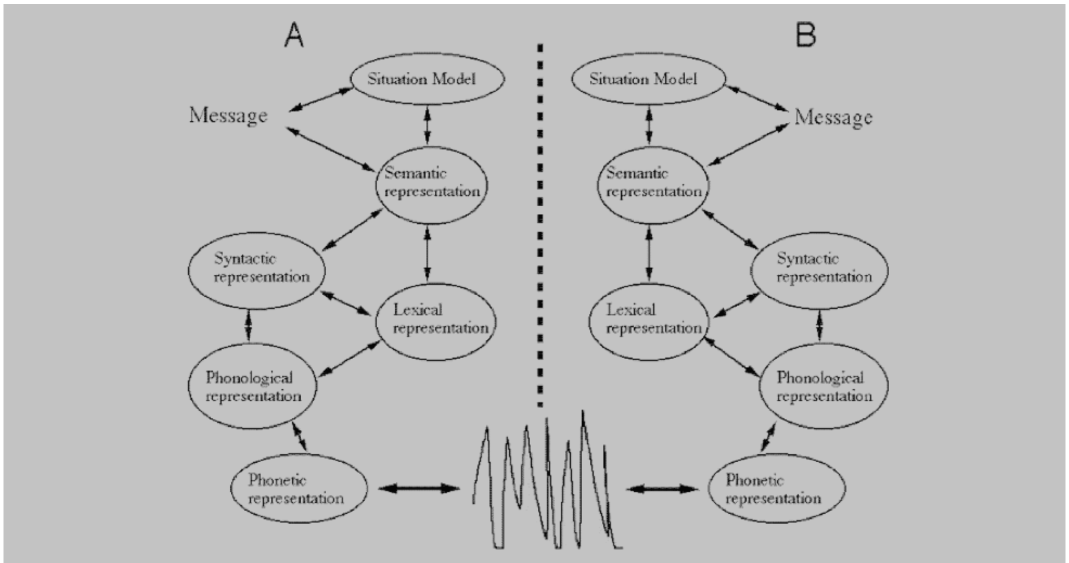


Figure 3. A and B represent two interlocutors in a dialogue in this schematic representation of the stages of comprehension and production processes according to the autonomous transmission account. The details of the various levels of representation and interactions between levels are chosen to illustrate the overall architecture of the system rather than to reflect commitment to a specific model.

**Figure 5.1. A schematic representation of the stages of comprehension and production processes of language (from Pickering and Garrod 2004)**

The diagram in Figure 5.1 above demonstrates this interconnectedness of various stages of the comprehension and production of language. The speaker or writer converts a non-linguistic "idea" or message into a series of linguistic representations, which are decoded by listeners and readers back into the message. By disambiguating and sequencing the various types of linguistic representation at play in the process of transmitting a message, we are better able to identify the discrete areas in which alignment may take place—for example, changing one's accent to better match that of a conversational partner could be considered one type of accommodation that is qualitatively different from employing a partner's word choice to describe a phenomenon. This model also illustrates how the relationship between language, cognition, and meaning is mutual and reflexive, with semantic choices influencing the situation model and interpretation of a message and vice versa. Indeed, this makes intuitive sense: extensive research on framing (Nelson et. al. 1997, Levin et. al. 1998) has long established that the framing used when communicating about an issue can drastically change perceptions and beliefs about reality (and can in the speaker reflect those perceptions and beliefs), and that ideologically similar individuals tend to use ideologically similar frames (Lakoff 2008).

It stands to reason that the semantic construction of these frames would leave lexical and syntactic footprints—talking about an issue from a conservative viewpoint may necessarily use different words (and, perhaps, different sentence structures) than would talking about the same issue from a liberal viewpoint. Though the literature on alignment has focused heavily on the cognitive processes involved in dialogue, the alignment model allows us to re-approach some of the broader social questions posed by the original discoverers of linguistic accommodation—for example, whether or not there are mechanisms beyond interpersonal rapport that result in style modulation. Armed with these theoretical tools, we may more broadly look at the phenomenon of convergence of various observable and measurable elements of dialogue. It is for the applicability of the methods researchers have used to analyze accommodation to blog corpora that we draw on this literature, situating it within our understanding of the interplay between situation model and linguistic representation.

It is important, though, to examine more closely the appropriateness of this framework to examine online blogging. As this theory was developed around the study of collocated, spoken conversation, some characteristics of the blog medium cause critical differences in our application of accommodation theory from the original context in which it was developed—the one-to-many nature of the blog medium, the lack of instantaneous feedback between speaker and audience, and the computer-mediated nature of the communication taking place. We address these very important contextual differences by examining the assumptions behind linguistic alignment and linguistic accommodation and by further examining contemporary use of these theories in empirical research.

The first difference we point out is that of the broadcast nature of blog authorship, as contrasted with the single-speaker single-audience-member nature of dyadic dialogue. The authors of linguistic alignment themselves point out that while alignment is primarily descriptive of the process of dialogue, monologue (for example, as it takes place in blogs or newspapers) can be considered a special, asymmetric case (one speaker, the author, speaks much more than the others, the commenters and readers). The author of a monologue (in our case, a blogger) is writing to an audience she is both aware of and from whom she often gets significant, rich feedback (Walker 2006, Dennen and Pashnyak 2008). Linguistic alignment has been observed in online communities where the primary mode of communication is not dialogue (Martin 2004) and even with nonhuman actors such as computer programs (Branigan 2010). Though dialogue in the conventional sense does not always occur between the authors of blogs (and other broadcast media) and the blog readership, some cohesiveness between the blog author's situation model and that of her readership is necessary in any successful communication of the author's viewpoint.

The second difference we need to address is that of synchronicity—in the case of an in-person conversation, alignment depends on instantaneously negotiated, multidimensional communicative feedback. Blog authors do not experience the same kind of feedback an interlocutor might, but evidence suggests that feedback does take place between a blog's author and his audience, despite the temporal and physical gulfs that may lie between them. It is true that the vast majority of a blog's audience does not write back, despite the fact that most blogging platforms allow for comments. While blog posts could not in themselves be considered directly part of a dialogue due to their broadcast nature, they are indisputably part of *discourse* with their readership, with authors of other textual media, and with other bloggers, categories that may not have defined boundaries and that are not mutually exclusive.

The social world of blogs is rich and it would be absurd to posit that bloggers are unmindful of their readers. Temporal gulfs notwithstanding, it is clear that bloggers and their audience are communing in complex and meaningful ways. Bloggers themselves are both consumers of mainstream media and producers of alternative media, in which bloggers and commenters often actively participate in discursive interpretation of events, as revealed by both ethnographic study and textual analysis (Mitra 2010, Boicu 2011).

Indeed, more specifically, alignment has also been observed between blog authors and their commenters (Goode and Robinson 2013), and in complex, asynchronously communicating communities (Martin 2004). It has been noted that political blogs in particular are "spaces where those of similar ideology [converse]," and that a significant amount of interaction takes place on these blog sites, with a great deal of "community information sharing" having been observed in comments (Walker 2007). Given the overwhelming evidence that political blogs function as rich, discursive communities, and given that linguistic alignment has been observed in these types of settings, it seems plausible that we might observe instances of alignment in the liberal and conservative corpora.

As the vast majority of studies involving alignment deal with in-person verbal communication, we must address whether computer mediation interferes with linguistic alignment. As previously discussed, linguistic alignment involves multiple dimensions of communication and the question has been raised as to whether it is meaningful to discuss alignment in a computer-mediated setting. Though theories about alignment and accommodation both emerged from studying in-person interactions, evidence has accumulated that accommodation can occur in online, computer-mediated interactions. In particular, experimental studies of intercultural collaboration (cf Wang and Fussell 2010) show not only that alignment occurs, but that it occurs even in electronically mediated, cross-cultural settings. Another observational study of alignment involved examining many conversations taking place on the Twitter microblogging platform (Danescu-Niculescu-Mizil et. al. 2011), which found that despite the radical differences between synchronous, collocated spoken conversation and asynchronous, computer-mediated written and character-limited conversation on Twitter, linguistic style accommodation as expressed through similar function word use was observable in pairs of people that tweeted to one another, even if those tweets weren't identified as belonging to temporally and/or narratively coherent "conversation". Similar to blogs, Twitter is an asynchronous, one-to-many broadcasting medium. The examples provided above of linguistic accommodation and alignment taking place in computer-mediated, non-synchronous and non-dyadic-pair settings support our assumption that linguistic alignment can be observable in computer-mediated settings.

The convergence in communication patterns originally observed in person by the progenitors of accommodation theory took place along a variety of dimensions, including dimensions not communicable through an online textual medium (e.g. gesticulations). It has been noted, though, that convergence in style did not necessarily occur along every possible dimension in every case. Examination of just the syntactic dimensions of accommodation, as is possible through LIWC, has proven to be very fruitful for researchers, who have found that accommodation evidenced by structural convergence is correlated with relationship success (Ireland et. al. 2011), productivity at work (Gonzales et. al. 2009), and negotiation outcomes (Taylor and Thomas 2008; Rogan 2011). While we could not, for example, directly observe the behaviors of the multitude of blog authors whose blogs we collected for this study, we can analyze the linguistic elements of their speech. It is, as such, appropriate in our study of blogs to focus on only the syntactic characteristics to observe accommodation. To do so, we employ the Linguistic Style

Matching (LSM) metrics developed by Pennebaker et. al, which we describe in more detail in a section below.

### *5.2.1 Political Blogs*

It is both interesting and appropriate to investigate the presence or absence of linguistic accommodation in the setting of communities of bloggers. A debate continues among academics as to the role of political blogs (and the Internet at large) in the political process. The Internet has long been at the focal point of claims about its effect on society and public society, with some, such as Benkler (2006) and Woodly (2008) arguing that blogs enhance democracy by allowing people to argue and engage with others with divergent viewpoints, largely free of coercive forces and control by governments. This position has found some empirical support (e..g Nahon and Hemsley 2011). Others (e.g. Sunstein 2003) predicted fragmentation and balkanization of information, with people self-selecting to receive information that aligns with their preexisting views.

A considerable amount of effort has been expended since then to determine which narrative, if either, is in fact correct. Approaches to this issue have been mixed. A number of researchers (e.g. Adamic 2005, Hargittai et. al. 2008) employed network analysis of blog links to find that liberal and conservative blogs each constitute very separate communities that rarely link to one another. The divide between liberals and conservatives is not limited only to (the lack of) cross-linking: Shaw and Benkler (2012) show that liberals and conservatives may prosecute the dissemination of information differently, exhibiting different behaviors regarding the origin, purpose, and structure of online discourse even within the singular medium of the blog.

In the context of the American political blogosphere, even those that believe networked communications lead to cross-ideological exchanges have acknowledged that the liberal and conservative blogospheres are distinct communities. As we have noted, linguistic accommodation takes place within online communities (Goode and Robinson 2013, Martin 2004), leading us to surmise that if accommodation were occurring among the blogs we studied, it would be taking place independently in the liberal political blogosphere and the conservative political blogosphere. If linguistic convergence were separately observed between liberal blog authors and between conservative blog authors, it would have a number of interesting implications both for our understanding of the impact of political ideology on society, for the applicability of linguistic accommodation theory to broadcast media, and for our understanding of blogging communities.

We hold, as others (Taylor and Thomas 2008, Niederhoffer and Pennebaker 2002, Pennebaker and King 1999) have, that linguistic style offers a sensitive and more effective way to measure the capacity for engaging with others, and that this capacity is largely driven by the propensity of interlocutors to find a mutually agreed framing of a given situation—a shared situation model. Finding significant differences in measures of linguistic style between liberal and conservative bloggers would be linguistic evidence that liberals and conservatives approach issues with different frames and different situation models.

## 5.3 Linguistic Style Matching

Style, as previously established, consists of linguistic factors orthogonal to the content of speech—factors which, when we attempt to study them in an empirical manner, can be examined through the use of non-content words by speakers—pronouns, prepositions, articles, and so forth. Examining the use of these types of words gives insight into the structural makeup of language. It is to examine this concept of style that Niederhoffer and Pennebaker (2002) developed Linguistic Style Matching (LSM), a measure formed by measuring the prevalence of non-content words, which include the following terms shown in Table 5.1:

| Type | Example |
|---|---|
| Article | A, an, the |
| Conjunction | Also, and, but |
| Impersonal pronoun | It, someone, anyone |
| Negation | Never, not |
| Preposition | Over, through, with |
| Quantifier | Many, fewer |
| Auxiliary Verbs | Go, will, help |
| Personal Pronoun | You, I, she |
| Adverbs | Quickly, reluctantly, brightly |

**Table 5.1. Categories of words in LIWC that are part of linguistic style (Niederhoffer and Pennebaker 2002)**

Niederhoffer and Pennebaker employed LSM by computing the proportional prevalence of each of these types of measures in the texts produced by indviduals or groups, and examining the mean prevalence of each metric in each different population for statistically significant differences.

Researchers have previously used LSM to predict a variety of behaviors—a study examined relationship outcomes, looking at the stability and long-term viability of romantic relationships as predicted by the prevalence of non-content words in their text messages. Couples who used a similar style were found to be significantly more viable in the long run (Ireland et. al. 2011). LSM has also been used to predict outcomes of group interaction in team settings (Fischer et. al. 2007, Gonzalez et. al. 2009), in which high matching of non-content words was correlated to better team performance and satisfaction with the collaboration. Outcomes of negotiations where the level of LSM was higher were also found to be more favorable (Taylor and Thomas 2008). These results confirm that the use of style in language is not merely an aesthetic consideration, but that style is related to many social and cognitive factors including power, demographics, life experiences, emotion, and purpose.

Awareness of an audience and self-reflectivity as such has been described as integral to the formation of linguistic style (Bell 1984). As such, while LSM has largely been used in investigations of dyadic pairs and small groups, inevitably all writers, including broadcasters such as reporters or bloggers, write with an audience in mind. Evidence that reflexive alteration of linguistic style occurs outside of the conversational paradigm exists, though research in this vein has been sparse—similarities in style were found to be prevalent within identifiable online communities, with participants coalescing on stylistic commonalities over time (Nguyen and Rosé 2011).

It is because LSM is a computational method that is well-suited to the study of digital communication that we use this approach to investigating linguistic alignment. While style has also been studied through painstaking analysis of a small number of individual texts, close readings of tens of thousands of blog posts would be all but impossible. While this computational approach to studying style is not a replacement for close readings, LIWC's measures quantitatively capture a meaningful slice of what theorists refer to as linguistic style.

In this study, we compare the linguistic style of text written by liberals and text written by conservatives, looking for significant differences in the prevalence of the textual markers corresponding to linguistic style. As we discussed, linguistic alignment can take place within communities, and the American political blogosphere has oriented itself around two communities, one liberal and one conservative. While in the previous chapter we have established that speech patterns related to gender were also found in liberals and conservatives respectively, we now seek to establish whether patterns of function word use that have been connected to formal definitions of linguistic style present a difference across political ideologies. Doing so allows us to draw on the significant body of research related to linguistic style, its origins, and how it is transmitted.

If linguistic alignment—a convergence among multiple parties in the way they use language—is taking place (separately) within liberal and conservative communities respectively, prior research such as that we have detailed suggests that it would be revealed in the LIWC measures identified in Table 5.1. If we were to find that liberal bloggers had converged on a certain way of speaking and conservatives had converged on a statistically different way of speaking, we caution that this alone would not be proof that linguistic convergence would be the explanatory mechanism for this phenomenon. However, the evidence that such convergence has been observed on an individual level within asynchronous, conversationally disjointed computer-mediated environments such as Twitter (Danescu-Niculescu-Mizil et. al. 2011) and in online communities (Nguyen and Rosé 2011) suggests that characterizing such a pattern of behavior as linguistic alignment would be at least plausible.

## 5.4 Liberal and Conservative Linguistic Styles

We present the following hypothesis regarding liberal and conservative linguistic style:

H1: Liberal bloggers and conservative bloggers will manifest differences in linguistic style by displaying statistically significant differences in the frequency of use of the LIWC categories that make up the LSM measure described in Table 5.1.

Though it has been noted that individual linguistic style varies quite widely (Taylor and Thomas 2008 “Linguistic Style Matching and Negotiation Outcome”), finding a significant difference between the liberal and conservative linguistic styles would have a number of implications.

First and foremost, this result would provide additional empirical evidence supporting the idea that liberal and conservative blogs form very distinct communities. Communities, as previously noted, tend to converge linguistically in terms of style, and that we would find significant differences is suggestive that linguistic alignment has taken place separately in the liberal and conservative blogospheres respectively, or that some factor that mediates political ideology has also mediated linguistic style.

Socioeconomic, demographic, moral, psychological, and behavioral differences between liberals and conservatives have long been documented in academic and popular contexts. In addition, the liberal and conservative blogospheres have been considered distinct communities—numerous analyses have demonstrated that the diversity of opinions and political beliefs among the readership of individual blogs is low. Significantly more cross-linkages were found among blogs with similar political persuasions (Adamic and Glance 2005). Political blogs have been described in the media as “rumor mills” and “ideological lynch mobs” (Rall 2005).

Referring back to linguistic accommodation, it is possible that the cohesion in worldview within the separate liberal and conservative blogospheres, which has been described in prior studies using qualitative and network analysis methods, would lead to quantifiable differences in the linguistic characteristics of texts from different communities with differing ideologies, though this possibility remains unexamined in the literature. It is not unreasonable to hypothesize that we would see observable evidence of linguistic style matching among liberals and conservatives respectively. Since liberals are more likely to link to and read other liberal blogs, and because blog authors cater to their audiences in a variety of ways, including content (Gurzick et. al. 2006), and framing (Johnson et. al. 2007), we think it is likely we would find differences in style between blogs of different political persuasions.

Discovering such differences would be interesting because to date, few studies have tackled differences in language use between writers of different political persuasions. Studies of this type have examined latent differences in people of different political persuasions, including moral foundations (Graham et. al. 2009), have described the online behavior and community characteristics of adherents of various political parties (Adamic and Glance 2005, Dehghani et. al. 2011), and have explored the framing of issues (Yano et. al. 2010). However, to date such linguistic inquiries have been limited to issues surrounding semantic content, and no studies of the interplay between non-content words and political ideology have been conducted.  Finding such differences would have implications both for our understanding of these communities but possibly also for our understanding of political ideologies and the way we structure our language.

In addition, an empirical study in this vein would be revelatory in that few studies to date have examined non-content words in blog communities—the bulk of inquiry regarding linguistic style, including that involving social media, has dealt with individuals, dyad pairs, small groups, and the characteristics shared by a population of individuals (see Pennebaker 2011, Ireland et. al. 2011). With few published studies to date examining linguistic style in a community setting (cf Alpers et. al. 2005, Vambheim et. al. 2012), further empirical examination of this area is merited. In addition, as we seek to examine communities bifurcated by political ideology, we may draw on the extensive theory on politics available as an interpretive lens.

## 5.5 Methods

For this study, we used the same dataset as we did in the previous chapter and that is described in the Methods chapter of this dissertation. Posts from May 2012 thru December 2012 were obtained from the top 50 liberal and top 50 conservative political blogs as listed by Technorati, with over 80,000 posts ultimately collected and analyzed.

For analysis, we again used the LIWC tool. LIWC measures the proportion of words in any given document that belong to any of the dictionaries it contains. For example, LIWC has listings of pronouns, prepositions, words that are associated with cognitive processes, and so forth. For this study we used the measures defined as LSM, described in Table 5.1 (see section 5.3).

Our usage of LSM differs somewhat from that of previous studies. LSM was pioneered for the study of dyadic pairs of interlocutors in controlled experimental settings (cf Taylor and Thomas 2008). However, the measure has shown itself to be useful outside of the lab, as evinced by its evolving usage: later, dyadic pairs out of the lab (Ireland et. al. 2011) and documents such as essays and poetry that were not directly part of any interpersonal dialogue (Ireland and Pennebaker 2010). LSM was also measured in text aggregated from more than one individual, during a study of groups that were completing a task together (Gonzales et. al. 2009). Though the LSM metric has not yet been applied to large groups in the aggregate, one study on linguistic style accommodation in Twitter used LIWC metrics the researchers identified as pertaining to linguistic style (Danescu-Niculescu-Mizil et. al. 2011).

Since this study uses a dataset with different characteristics from that found in prior studies, we chose to use statistical tools well-suited to our particular needs—we are not measuring cohesion within a single population or a comparison between multiple dyadic pairs of individuals. GLMM, also used in the previous chapter, proved an adequate statistical tool for this purpose. As previously described, GLMM is a regression model which predicts the expected value of a response variable as a linear combination of a set of predictors. GLMM is a powerful statistical tool, because unlike standard linear regression, it does not assume that the predictor variables have a normal distribution, nor does it assume the independence of predictor variables. This last distinction is particularly important as it has long since been established that linguistic metrics cannot be assumed to be independent (Radday and Wickman 1975). Intuitively, this makes sense, since the rules of language dictate that some words must follow others and thus that use of some categories of words might be correlated with use of other categories.

With GLMM, different configurations of metrics used as predictors for a given model would, depending on the level of intercorrelation of a given combination of metrics, result in different correlations between predictors and response variables. Our use of GLMM focused on the examination of whether or not high or low values of the various LSM metrics were correlated with liberalism or conservatism. If liberals and conservatives use, on average, the same distribution of linguistic styles as one another, it would follow that significant differences in use of these style metrics would not be found.

## 5.6 Analysis

To test this, we created a statistical model using GLMM, where our target was a binary variable—whether the blog was liberal and conservative. The model describes how much variance is accounted for by the aforementioned LSM factors for liberal versus conservative blog posts. Again, we used blog ID as the random effects measure. By grouping the data by blog, we eliminate the possibility that one or more particularly verbose blog authors skewed the data. The results of the statistical model are displayed below:

| **Measure** | **Coefficient** | **Significance** | **Mean (lib.)** | **Std. (lib.)** | **Mean (con.)** | **Std. (con.)** |
|---|---|---|---|---|---|---|
| articles | 0.010 | .000 | 7.28 | 3.41 | 7.31 | 3.61 |
| auxiliary verbs | -0.013 | .000 | 6.83 | 3.82 | 6.09 | 3.26 |
| conjunctions | 0.009 | .020 | 4.62 | 2.75 | 4.41 | 2.45 |
| quantitative | -0.017 | .000 | 2.46 | 2.11 | 2.30 | 2.25 |
| prepositions | 0.040 | .000 | 12.67 | 4.80 | 13.21 | 4.39 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| adverbs | -0.022 | .000 | 3.31 | 2.91 | 2.89 | 3.05 |
| impersonal pronouns | -0.080 | .000 | 4.63 | 3.28 | 3.87 | 2.50 |
| personal pronouns | -0.049 | .000 | 4.42 | 3.63 | 3.68 | 2.98 |
| negations | -0.020 | .001 | 1.29 | 1.55 | 1.10 | 1.45 |

**Table 5.2. Output for Linguistic Style Model**

Table 5.2 above is a summary of the output of the GLMM. The model incorporated all of the measures comprising linguistic style matching—articles, auxiliary verbs, conjunctions, quantitative words, prepositions, adverbs, impersonal pronouns, personal pronouns, and negations. As was the case in the previous chapter, as the reference value for the binomial target variable (liberal vs. conservative) was liberal, positive coefficients meant the term was used more often by conservatives, and negative coefficients meant the term was more used more often by liberals. The results above are telling: liberal and conservative bloggers used significantly different quantities of these types of words, meaning that they differed significantly in their linguistic style.

We found that of the nine factors that make up LSM—articles, auxiliary verbs, conjunctions, quantitative words, prepositions, adverbs, impersonal pronouns, personal pronouns, and negations, all nine were statistically different between liberals and conservatives, with a significance of < 0.05. Conservatives used more conjunctions and prepositions, while liberals used the other types of words—articles, auxiliary verbs, quantitative words, adverbs, impersonal pronouns, personal pronouns, and negations—in greater proportion. We note that individually, some effect sizes were larger than the 0.10

value for Cohen’s *d* that characterizes a minimum notable effect size, while for other measures, values of *d* were smaller than 0.10. However, the statistical treatment of LSM in previous studies has always considered LSM metrics in the aggregate since many contextual factors affect which indivdiual play a greater or larger role in any given dataset. The average effect size (Cohen’s *d*) of the group of LSM metrics was 0.12, which is within the range of significance as established by previous studies using LIWC on similar data sets (Newman et. al. 2008).

The above results indicate that liberal and conservative blogs tend to have stylistic similarities within their ideological population, but statistically differ from one another in use of non-content words. Had there been too much variance in either population, we would not have found that these measures differed significantly between the two populations. We discovered differences despite the fact that the two populations are quite similar in other demographics that have been identified as having effects on style—having roughly the same gender balance, relatively affluent, and from the same country, for example.

We again felt it critical to test whether author gender had a mediating effect on the significant results we identified on style, and to test this we incorporated the masculinity measures developed in Chapter 4 as random effects into three separate GLMM models, using the same fixed effects terms as above. We report the significance of the fixed effects terms with various scales for blog gender included as random effects below, where each coefficient column details the fixed-effects coefficients for the model with the indicated scale for blogger gender incorporated as a random effect, accompanied by the significance of the fixed effect:

| **Measure** | **Coefficient w/ masculinity_75** | **Sig.** | **Coefficient w/ masculinity_50** | **Sig.** | **Coefficient w/ masculinity w/ unknowns removed** | **Sig.** |
|---|---|---|---|---|---|---|
| articles | 0.023 | .000** | 0.019 | .000** | 0.022 | .000** |
| auxiliary verbs | -0.027 | .000** | -0.002 | .000** | -0.020 | .000** |
| conjunctions | -0.007 | .129 | -0.023 | .659 | -0.010 | .000** |
| quantitative | -0.024 | .000** | -0.026 | .000** | -0.023 | .000** |
| prepositions | 0.028 | .000** | -0.021 | .000** | 0.026 | .000** |
| adverbs | -0.016 | .000** | -0.011 | .002** | -0.014 | .000** |
| impersonal pronouns | -0.111 | .000** | -0.122 | .000** | -0.133 | .000** |
| personal pronouns | -0.037 | .000** | -0.053 | .000** | -0.041 | .000** |
| negations | -0.016 | .040* | -0.002 | .762 | -0.012 | .107 |

**Table 5.3**. **Output for Linguistic Style Model with author gender as random effect.**

While for the most part, the tendency of liberal blogs to have different linguistic style from conservative blogs is maintained when we control for author gender, the analyses do show that gender does mediate language use to some degree. While all linguistic style terms displayed significant differences when blog gender was not controlled for, in each of the above models at least one term no longer displayed a significant correlation. Again, a positive coefficient indicates that the predictor was associated with conservative blog posts, where a negative coefficient indicates that the predictor was associated with liberal blog posts.

Use of conjunctions and use of negations in particular showed sensitivity to the inclusion of some of our author gender measures as random effects. It is possible that this is symptomatic purely of the small effect sizes we observed in many of these measures. In any case, author gender does not appear to explain enough of the variance in language use we observe between the liberal and conservative blogging populations to be explanatory of our findings.

## 5.7 Limitations

An often-observed difficulty of studying natural language is the immense amount of variation in how often words are used depending on the context. Numerous factors, including the subject of an individual text, dramatically affect the statistical properties of a given text. Similar studies of field data not obtained through the controlled environments of experimentation (cf Newman et. al. 2008) have noted the difficulty of examining language use because language varies so greatly depending on a wide variety of cross-cutting contexts. Though the effect sizes we observed were small, they were similar in size to those found in previous studies.

As previously mentioned, style changes contextually from individual to individual and also from text to text depending on audience, history, demographic factors, purpose, and probably the content of the text itself. As such, a statistical approach has its strengths because of our ability to wash out a decent proportion of these differences as random variability—however, careful consideration of the factors that may affect style is merited, and we cannot exclude the possibility that some external factor can explain the stark differences between liberals and conservatives that we found in the data.

We acknowledge that we were unable to reliably obtain from blogs important metadata—for example, a number of demographic factors are known to affect linguistic style, such as age. We did show earlier that the data set did not show evidence of differences in function word usage related to age or to social class. However, without direct demographic information of bloggers and, critically, a more complete understanding of what kinds of environmental and structural factors affect linguistic style, we can't rule out the possibility that some extraneous factor other than political ideology may be causing the differences we observed.

Another important limitation of this study is that the quantitative study of linguistic style is still nascent, and it is inevitable that other meaningful ways of describing or conceiving style have not yet been devised. Compared to the possible number of channels through which participants in a dialogue can interact, we focused only on word frequencies of the structural elements of language. Omitted from this study (and for the most part, from the quantitative study of linguistic style) is the analysis of syntactical and emotive representation in writing. While these elements are difficult to process computationally, doing so could add depth and reliability to analyses of style.

In this vein, it would be useful to bridge the gap between the qualitative and quantitative conceptualizations of linguistic style. While we used Niederhoffer and Pennebaker's

(2002) computational rendition of style, style encompasses many different overlapping social meanings that include aesthetics, phonology, indexicality, and stance. While exploratory studies such as ours can be revelatory, there is a need for theory to tie together the various understandings of textual data that have been informed by different methods situated in different disciplines. It is apparent that computerized textual research is still in an early stage, and as such we stress an interdisciplinary approach to solving the puzzle of how humans and our language are interrelated.

## 5.8 Discussion

That liberals and conservatives display clear differences in their use of linguistic style has a number of implications. While the difference in means was small, the distribution of effect sizes fell solidly within the range of those found in previous studies with similar data sets.

The implication is thus that the differences we observe in these two populations is connected to political ideology. Ideology is the filter through which we see the world as informed by our most basic moral and ethical beliefs. That political ideology might affect a writer's approach to the structure of language in addition to the well-documented differences in content has intriguing consequences; our finding seems to confirm that the theory of linguistic alignment and the idea that one's worldview may have effects on how one expresses one's thoughts at a basic, reflexive, and subliminal level.

Our finding that linguistic alignment takes place within communities organized around political ideology invokes questions about the nature of this alignment. There are a number of possible explanations. Are bloggers finding and imitating cues from one another or from their audience, or is there something innate about political ideology that affects the production of written text in such a way that stylistic differences are measurable? The precise origin of the linguistic differences between liberal and conservative remain an area of speculation. Though we now know that political ideology has an effect, to determine why and how it has an effect, further study will be required.

In addition, Niederhoffer and Pennebaker's prior findings that differences in linguistic style have a relationship to the success of interpersonal relations is both intriguing and disturbing. That the propensity of people to speak may similarly correlate to relationship success and negotiation outcomes is particularly concerning in the face of the striking differences in liberal and conservative use of language. If contemporary liberals and conservatives are in some ways *speaking a different language*, in the context of today's polarized politics, we might anticipate difficulties in the day-to-day business of getting along with one's neighbors and co-workers. Though it is important to emphasize that we are not positing a causal relationship between linguistic style and interpersonal success—the direction of influence is unclear and probably quite complex—it seems to be the case that different American political frames produce marked differences in the style of language production. Liberals and conservatives have long been popularly accused of being unable to communicate with one another about politics. Linguistic style could be one window into understanding why and how this might be the case.

Our result also corroborates prior findings (Adamic and Glance 2005, Hargittai et. al. 2008, Lawrence et. al. 2010) that liberal and conservative blog authors are writing for different audiences, likely audiences with similar beliefs to the writer. This is one more nail in the coffin for the idea that the blogosphere might provide a marketplace of ideas where people of various beliefs find viewpoints different from their own. Rather, the stylistic differences are suggestive of a milieu in which liberals and conservatives are self-segregated into discrete communities.

We find further confirmation that linguistic style is relevant to the study of communities. While further study is necessary to fully explore the relationship of style to different kinds of communities, it does pose a number of interesting questions. What causes various kinds of communities to differ in style? Do all communities show evidence of style matching, and if not, what types do? Do communities not formed around an ideological subject matter differ from one another after effects of demography are controlled for?

Interlocutors who have similar styles tend to have higher outcomes of relationship success (Ireland et. al. 2011) and teamwork (Gonzales et. al. 2009). Indeed, Ireland et al. found that 76.7 percent of couples displaying LSM cohesion above the median (based on a single speed dating transcript) were still together when researchers followed up with them three months later. Only 53.5% of couples with LSM cohesion at or below the median were still together. It has further been theorized that matching LSM indicates a matching in information organization processes and problem-solving processes (Tausczik and Pennebaker 2013). That these effects were found to be so dramatic has implications for the way we might think about our results: it is possible that liberals and conservatives might have a lower chance of getting along in relationships, or working with one another in teams, whether or not they ever speak about politics with one another. Studies have not yet been conducted to determine whether liberals and conservatives are able to accommodate after speaking to one another—the starting point of having differing linguistic styles may make it harder for them at the outset, and since linguistic style is so unconscious, it seems reasonable to theorize that liberal and conservative differences go much deeper than simply having differing opinions. However, linguistic accommodation is by all accounts a reflex that has been observed across many contexts. What happens when liberals and conservatives communicate with one another is a ripe area for future inquiry.

When Lakoff (2010) says that "conservatives like to make fun of liberals, claiming that liberals just don't speak their language," he may be literally correct. "Big government," he says, does not just refer to a large government—it is laden with meaning particular to the conservative lexicon. He offers that liberal rebuttals to accusations of "big government" that refer to prison and military spending are often laughed off, because "the liberals have just misused the term." Though Lakoff was referring only to lexical differences, our finding indicates that syntactic, structural differences are also apparent. The contrast we found in the prevalence of non-content words indicates that between liberals and conservatives, the very fabric of the language that is being used by these different groups has discernible differences.

A puzzle that remains for future researchers is the question of what these metrics might mean, both individually and in the aggregate. We turn to the sociolinguistic model of communication shown in Figure 5.1 to examine the possibilities. First, it is possible that the syntactic differences detected are purely associated with and explainable by the pattern of lexical differences already detailed by many academics, that terms like "big government" are simply naturally associated with fewer pronouns and more articles than the term "safety net" as a consequence of the relationship between lexicon and syntax that has been programmed into the English language. The model shows that there is a bidirectional relationship between lexical representation and syntactic representation, in that what words one chooses may affect what syntax one may need to use and vice versa—passive constructions require different verbs from active constructions.

More intriguing, though, is another possibility that is validated by the sociolinguistic communication model—that the syntactic differences we observed reflect differences in state of mind, both in the way people understand situations and communicate ideas, as Pennebaker and others have proposed. An individual's situation model, and the text and subtext they wish to communicate verbally as a result, has impacts on *both* lexicon and syntax.

Though many fascinating studies exist on how observable differences in style are correlated with observable differences in individual behavior and outcomes, it is difficult to assign meaning to the relative prevalence of any individual function word. Typically, quantitative analysis of language has involved metrics with clearer connections to intuitively understood phenomena, as is the case with sentiment analysis. A generalizable understanding of the use of non-content words remains elusive. However, there does not exist a definitive explanation of what exactly it would mean for one individual to use (for example) more pronouns than another. We offer that this line of inquiry may not be the most useful—rather, we suspect that it would be more fruitful to continue to examine the relationship between various aggregations of function word use and observable psychology and behavior.

# Chapter 6 : Time Series Analysis of LIWC Measures

## 6.1 Introduction

The previous chapters have dealt with the examination of differences in the use of linguistic markers among the liberal and conservative blogging populations as aggregated over a period of time. In this chapter, we continue this comparison of the use of language between these two populations with a sensitivity to changes over time. Most studies involving LIWC have noted how noisy data from real-world natural language sources can be, and on the whole they have taken a number of measures to mitigate noise in order to obtain more meaningful results, including collapsing multiple data series within subjects (Pennebaker et. al. 2004), collapsing (Mishne and Glance 2006) or smoothing (O'Connor et. al. 2010) data across time periods, or they have simply presented a case that small effect sizes remain meaningful (Newman et. al. 2008). Taking a bottom-up, exploratory approach, we examine assumptions and claims about the behavior and meaning of style word use while seeking to better understand liberal and conservative blogs.

### *6.1.1 Analysis of literary style*

A longstanding assumption of *stylometry*—that is, the use of statistical methods in the analysis of literary style (Holmes 1998, Koppel and Schler 2003) is that stylistic metrics such as the ones used by LIWC should be relatively invariant within the works of a given author but vary from author to author depending on characteristics of that author. That an author might not change her writing style as observed through style metrics from work to work has been analytically validated (Holmes 1998) and achieving this invariance has been a goal in the development of many such style metrics—it is seen as a sign that the metric is sound, as the assumption of invariance in the metric is actually serving as a proxy for an assumption of invariance in style itself—which, as we have established in previous chapters, is a concept that is difficult to pin down.

The contexts of textual production as public record have changed significantly. Most applications of stylometrics usually involved the examination of books (cf Holmes 1992, Matthews and Merriam 1993, Merriam and Matthews 1994) and occasionally other texts such as periodicals (Cortina-Borja and Chappas 2006) and music lyrics (Whissell 1996). In the above examples, style was used as a tool for attribution and classification of texts and authorship—for example, Cortina-Borja and Chappas attempted to classify texts by medium—broadsheet, tabloid, etc.—using quantitative style metrics. Also, researchers used stylometric methods to distinguish between songs written by John Lennon and Paul McCartney, with Lennon's lyrics being sadder and less pleasant than McCartney's.

Other studies have used style metrics in an experimental fashion to approach social psychology problems, such as predicting deception (Newman et. al. 2003). The common thread in these studies is that researchers have generally appropriated style metrics for use in categorization and classification, as we have in the prior two chapters of this dissertation. This is unsurprising, as the textual data analyzed lends itself well to this manner of inquiry.

As we discussed at some length in the previous chapter, using linguistic style metrics to examine digital texts seems to be an obvious application of this class of methods. However, there are critical implications for the use of these methods in digital texts that we feel must be explored. The context of use of stylometry has evolved: earlier studies exclusively (and many contemporary studies still) dealt with formally published works, which by their nature had a much higher word count per document, had fewer publications per author, and fewer documents to analyze in general even across a lifetime. Studies of textual production within social media, on the other hand, can easily involve many more authors, many more documents per author, and far fewer words per document. These differences, already quite distinct in blogs, are even more distinct in microblogs, in which character-limited posts are quite short but are published much more frequently than more verbose media.

The prototypical use of stylometric methods is the analysis of the book, a genre that has significant structural differences from that of blogs. By expanding the use of stylometric research to the blog medium, we expand the range of possible research questions. However, the change in context involved with research on blogs necessitates the examination of methodological assumptions. First, books, which often take months or years to complete, are not usually completed strictly sequentially from front to back. On the other hand, blog posts, being much shorter, are usually written over a period of minutes or hours. The fact that blogs are shorter in length, more numerous, and time stamped means that we can understand how the blog is temporally situated with a much finer granularity than we can with books. Also, that the average book is orders of magnitude longer than the average blog post might explain the finding that relatively little variance in function word use was found from text to text for any given book author (cf Holmes 1998), while enormous variance in function word use is found from text to text in blog media (cf Chung and Pennebaker 2007). These results are critical in informing our approach—we are aggregating the textual production of many blog authors together, treating liberals and conservatives respectively as populations. The large variance in function word use from text to text necessitates our examination of a large number of texts, more than that available from a single source.

### *6.1.2 Style and temporal factors*

The invariance of style has been both demonstrated in studies of books (cf Holmes 1998) and is a common assumption in contemporary studies of linguistic style (Pennebaker 2011). However, we have established that the medium being studied is tacitly assumed to be part of the context of this claim. Does this assumption indeed break down when style is assessed on a short-term basis? In what way, if any, do measurements of linguistic style change on a short-term basis? Some have already experimentally approached the issue of stability of style across time and context. Gleser (1959) asked study participants to narrate their life experiences for five minutes, comparing categories of words between the two halves of their narration, finding a correlation of 0.5 on average across the various metrics. Others, such as Schnurr et. al. (1986) and Mehl and Pennebaker (2003) sample conversations days or and weeks apart, respectively, again finding a high degree of within-person correlation. Pennebaker and King (1999) examined a large number of text samples taken from diaries and college assignments written over a period of years, again finding that across time, medium, and topic, individuals were very internally consistent in their use of words.

The assumption of relative invariance in linguistic style as shown through function word use has already been called into question: there is support for the idea that use of function words can change over time in certain situations. The speeches of Rudy Giuliani were examined over the course of his career, and in particular, his use of "I-words" (I, me, myself) and "We-words" (we, us, ourselves). His speech was found to have changed significantly by the end of his mayoralty, influenced by personal upheavals and changing political fortunes. Cited as particularly influential were feelings of uncertainty and significant traumas (such as the 9/11 attacks) experienced over his tenure (Pennebaker and Lay 2002). Giuliani's words were compared against those of Shakespeare's beleaguered King Lear, whose pronoun use changed over time in similar ways as his political situation collapsed (Pennebaker 2011). Real political figures were compared with one another, as well: the interviews, press conferences, and debates of a number of presidents and presidential candidates were analyzed, and the structure of their writing was found to differ greatly depending on the challenges of their presidencies (Slatcher et. al. 2007). The style of presidential inaugural addresses were found to vary based on economic and other factors of the times (Whissell and Sigelman 2001). Another paper on I-word and we-word analysis dealt with verbal responses to various tragedies and traumas, describing how tragedies tend to bring people together and foster a sense of community. During traumas, we-words were found to have been used more often and I-words were found to have been used less often (Mark et. al. 2012).

In addition, though invariance of style was demonstrated through within-subjects correlation as summarized above, we could find no studies confirming style invariance that examined the behavior over time of style metrics applied to groups or communities, nor did we find evidence of style invariance in longitudinal studies involving a large number of daily measurements. We find it appropriate to once again examine the assumption of invariance and more broadly, what changes in style words look like over time.

Though we have established the plausibility that time and changing circumstances might affect the way one uses function words, these studies have focused on behavior or characteristics observed over a single, fixed time period across many individuals (e.g. gender, age, power, negotiation success—see Pennebaker 2011), or on changes in an individual over long periods of time (e.g. Shakespeare's writing, the rise and fall of Rudy Giuliani's political career). Few studies to date have examined trends found in time series data to examine the issue of how and why use of function words within groups and communities of people might change on a shorter time scale, instead focusing on how changing conditions on a lifetime, years- or decades-long scale have affected the use of language.

Until recent years, the use of data sets orders of magnitude smaller than ours was the norm in linguistic studies. As such, there have been relatively few studies of function words and linguistic style that have used time series, all of which have involved sentiment analysis. Gilbert and Karahalios (2010) described correlations in stock market behavior and sentiment in blog posts. O'Connor et. al. (2010) found correlations between blog sentiment and public opinion time series such as political polling and consumer sentiment. The recent studies of this type have involved millions or even billions of blog or microblog entries, and for the most part have been conducted within departments of Computer Science, likely due to the technical difficulty of obtaining and analyzing data sets of this magnitude. Large data sets are helpful to the study of time series, as segmenting blog data sets by day reduces the number of data entries per day by a factor of, on average, the length of the period from which data has been collected. For a study spanning half a year, this is a reduction of two orders of magnitude on a per-day basis. This has the potential to introduce additional variance and "noise" that can interfere with analyses.

## 6.2 Research Questions and Approach

There is a gap in our understanding of how function words behave—namely, how their use fluctuates in the short term. As it stands, there does not exist an understanding of how or whether shorter-time-scale influences might have an effect on the structure of language, as seen through the use of function words. Though the noise inherent in natural language data from the field is, as one researcher described it, a source of "frustration," (Chung and Pennebaker 2007), it has not been determined whether the short-term fluctuations in uses of function words are truly indecipherable. Though much research on style has focused on the tendency of style to be influenced by more permanent factors such as personality (Pennebaker et. al. 2001) and gender (Chung and Pennebaker 2007), the presence or absence of temporary factors such as trauma and stress (Mark et. al. 2012, Pennebaker 1993) have been shown to be correlated to stylistic changes as well. As such, we ask the following questions:

Is there evidence of short-term change in reaction to environmental factors in the use of function words and affective language in blogs, and if so, is there any discernible pattern to these changes? To what extent can we account for and begin to explain this seemingly random behavior? Can influences in the short-term, whatever those might be, result in differences in the expression of function words, affective words, or other style metrics?

Our approach to answering these questions begins with the unique structure of our data set. Our data consists of natural language text drawn from two distinct populations, American liberal political blogs and American conservative political blogs. Liberal and conservative blogs belong to separate communities that both deal with the same fundamental subject matter—American politics—but paradoxically interact with each other only minimally (Adamic and Glance 2005, Newman 2006). Both attract and are catered towards politically active readers, but each community has a distinctly ideologically polarized leadership (Lawrence et. al. 2010). Though the way they talk about and frame topics is ideologically driven and vastly different (Xenos 2010, Entman

2010), liberal and conservative blogs have been shown to both be driven topically to a large extent by the mass media (Wallsten 2007). It would not be unreasonable to surmise that because our data set was taken from the period around a U.S. Presidential election, both liberals and conservatives dealt with the election and its vicissitudes as a topic of discussion. Though not essential to our argumentation, a topic modeling analysis not included in this dissertation revealed that nearly half of blog posts over our time period referenced the election.

As explained throughout this dissertation, there are many factors that have been revealed to be more or less correlated with the output of linguistic style—in an earlier section, we compared linguistic style markers to fingerprints—detectable traces that are necessarily left by any verbal activity. One of the most effective research strategies for understanding differences in linguistic style, which is affected by the myriad factors mentioned above, is to isolate a given factor and analyze the data around a given manipulation. For the most part, we have isolated a single factor—political ideology, and examined the pattern of differences for insight into differences between the psychology of liberalism and conservatism. With attention to the unique nature of the population being studied, we also wonder if variation in liberals and conservative language in the short term might be observed as having a positive or negative correlation. Such a correlation would suggest that some detectable external stimulus or set of stimuli is having an effect on both liberal and conservative populations as a whole. If stylistic variation is not simply the result of the random collection of contextual factors in which each blog text is situated, but is instead systematic and experienced by a population as a whole, it should be reflected in a correlation, positive or negative, of the frequency of use of function words by these separate populations.

### *6.2.1 Approach*

A relatively simple test can be devised to examine these questions using those measures available to us. The tool LIWC, created by Pennebaker et. al. (2007), measures the prevalence of various parts of speech and types of words by comparing words in documents against dictionaries of those types of words. Notably, LIWC allows researchers to examine structural characteristics of language use by measuring the prevalence of different kinds of words that have been coded as function words, cognitive words, emotional words, social words, or personal concerns.

We can use this tool to examine textual data from our dataset, which consists of posts from liberal and conservative bloggers around the time of the 2012 US Presidential election. These are two distinct populations, which are separate communities and have different worldviews as described in previous chapters.

Examination by time series offers us another opportunity to discover another way in which liberals and conservatives are different—in their reaction to short term stimuli, perhaps—or, perhaps, we will find that liberals and conservative psychology is not entirely different. If we compare time series data on style markers within these two discrete populations, there are three possibilities: positive cross-correlation, negative-cross correlation, and no significant correlation. Understanding *if* and *how* the use of function words changes over time has a number of implications, both for our understanding of ideology and for our methodological understanding of function words.

If there is a negative correlation in how the two populations express the various style metrics over time, we learn that to some extent, these short-term fluctuations in linguistic style that many have termed “noise” are likely the result of some stimulus to which liberals and conservatives are reacting in opposite ways. We might investigate whether the well-known differences in liberal and conservative framing of the various issues under discussion might be leaving the “fingerprints” we observe. There may be implications for our methods in this case, as this would merit an investigation into what might be causing those short-term fluctuations. If fluctuations can be identified as systematic, they may be able to be controlled for, allowing better analysis of linguistic style in the future.

If there is a positive correlation in how liberals and conservatives use blog data, this suggest that there is some external factor or sets of factors to which liberals and conservatives are reacting *similarly*. This result would be juxtaposed against the results of the previous chapters, which demonstrate how liberal and conservative language differs and argue that these differences are the result of some underlying psychology. We would be drawn to explore what aspects of liberal and conservative psychology might share commonalities, though this result introduces complications: as we have controlled only for political ideology, a negative correlation would point to political ideology as the primary explanatory factor in the result. However, in the case of a positive correlation in the use of liberal and conservative use of function words over time, we will have only *eliminated* political ideology as an explanation for these short-term fluctuations. Also, as with the case of a negative correlation, there would be implications for methods—systematically positive correlations between style metrics over time might suggest that some external factor or set of factors affecting linguistic markers is both detectable and may be able to be controlled for.

The last possible result is perhaps the least interesting—no pattern of correlations between liberal and conservative use of function words over time would mean that we are no closer to finding an adequate explanation for fluctuations in function word use on a short-term basis. Such a result would further confirm the complex, highly contextual nature of natural language texts from the wild.

In exploring these questions, we see how words vary across a large sample of text, measuring the prevalence of linguistic elements including function words, words associated with cognitive mechanisms, and words associated with personal needs. We take this wide-net approach in part because it is the same approach that has been used to validate prior findings that style and its associated metrics are invariant within authors (Holmes 1998), and because it allows us to better examine broad patterns of language use.

By comparing time series data obtained using LIWC with our corpus of blog posts, observing the presence, absence, and direction of cross-correlations between the data from liberal and conservative blog texts will allow us to evaluate the aforementioned research questions.

We can speculate on any number of factors that operate on the time scale of days that might affect the psychological state in such a way that the structure of language would change. In looking at our chosen population, political blogs during an American presidential election, the projected electoral outcome for the blogger's favored candidate might produce more or less anxiety as the likely result is more or less in conformation with the writer's preference. Other studies have suggested that there is a relationship between polling numbers and political blog content (O'Connor et. al. 2010). Examples of external stimuli that affect blog content are the news cycle, which studies have shown drives blog content (Meraz 2009, Meraz 2011) and may cause both liberal and conservative blogs to be writing on the same topic at the same time.

## 6.3 Methods

We examined cross-correlations between time series data of prevalence of categories of words obtained through LIWC in liberal and conservative blogs. We examined all non-content categories of words analyzable through LIWC, which covered the gamut of functional words and verb tenses. Table 6.1 enumerates those metrics:

| **Category** | **Example Words** |
|---|---|
| Pronouns | I, you, it, everyone |
| Personal pronouns | I, you, we, she |
| I-words | I, me, my |
| We-words | We, us, our |
| You-words | You, your, yours |
| She/he | She, her, he, his |
| They | They, their |
| Impersonal pronouns | Everyone, anyone, it |
| Articles | The, a, an |
| Verbs | Go, speak, think |
| Auxiliary verbs | Can, should, must |
| Past tense | Ran, asked, found |
| Present tense | Run, ask, find |
| Future tense | Will run, shall ask |
| Adverbs | Quickly, thoroughly |
| Prepositions | Over, through, beyond |
| Conjunctions | And, or, but |
| Negations | No, not, never |
| Quantitative words | How much, how many |
| Numbers | 1, two thousand, thirty-two |
| Assent | Yes, Okay, Agree |

**Table 6.1. List of non-content linguistic markers available for analysis in LIWC.**

Since many of the prior studies involving time series also involved sentiment analysis, we also examined sentiment in our dataset. Sentiment, as described later on, was calculated using the relative prevalence of positive and negative words.

Using LIWC, all 81,649 posts in our dataset were treated as one document each, and LIWC generated counts of each of the above measures. To generate the time series, the data was aggregated by finding the average values of the LIWC measures for all posts on a given day within our range of dates. This was performed separately for liberal and conservative blog posts. The average number of posts per day was 177.5 in the liberal dataset and 155.6 in the conservative dataset.

### *6.3.1 Data Smoothing*

Temporal *data smoothing* has been found to be an effective strategy in deriving meaningful results from time series NLP data (O'Connor et. al. 2010). It smooths out short-term fluctuations and makes clearer the long-term trends in the data. Consistent with that study's recommendation, we have found that the basic and very commonly used smoothing technique of the moving average to be extremely effective. A moving average of a time series *x*, which we will call *MA,* with a window of size *n* for a given time *t* is calculated as follows, where $x_t$ is the value of the time series at time *t*:

$$MA_n(t) = \frac{x_t + x_{t-1} + x_{t-2} + \cdots + x_{t-n+1}}{n}$$

A moving average lessens the impact of high-frequency day-to-day changes in a metric by combining each day's value with that of *n* previous days. As an example, if *x* is the value of a stock, *t* is the day, and *n* is the window, which we will for the purposes of this example say is 7 days long, the moving average for a given day is calculated by adding the value of the stock for that day with its value on the previous six days, and dividing the result by 7. The threshold of smoothing between short-term and long-term can vary—the use of a long threshold can obscure some detail in the temporal fluctuations of a data series.

### *6.3.2 Cross-Correlation Functions*

With discrete time series data such as that produced by our data set, *correlation* refers to the degree with which two time-series vectors tend to deviate from their expected values in similar ways. A statistically significant correlation would mean that two time series tend to vary with one another. The formula describing the correlation between two time series *x* and *y* at lag *m* is shown below:

$$\rho_{x,y}(m) = \frac{\sum\left[(x_n - \mu_x)(y_{n+m} - \mu_y)\right]}{(\sigma_x \sigma_y)}$$

The signal expressed by two discrete functions can be analyzed for *cross-correlations,* which computes the changes in correlation between two time series as the time index of one of the two time series is sequentially manipulated. So rather than just comparing *x* at time *t* with *y* at time *t*, *x* at time *t* might be compared with *y* at time *t*+1, *t*+2 … *t*+*m*. The incremented unit of time by which one of the series is offset is referred to as *lag.*

Cross-correlation functions allow researchers to determine whether two data sets are just correlated, tending to occur at the same time, or whether statistical behavior in one time series anticipates behavior in another. Correlation of time series found at nonzero lag values can be considered a possible signal of a causative relationship, though this is not always the case.

## 6.4 Results

We examined the aforementioned time series of the LIWC measures listed in the Methods section, comparing use of every category of words among liberals with use of these categories among conservatives, over time. Seven-day moving averages were calculated and plotted. We used a seven-day window for our temporal smoothing as it was shown to be effective in prior NLP studies involving smoothing for improving correlations of linguistic data with real-world behavior such as polls (O'Connor et. al. 2010). Though O'Connor et. al. also identified longer windows as being effective, the seven-day moving average is a superior option for our study because the shorter window helps to mitigate the possibility of producing false correlations in our temporally smoothed data sets, an issue we explore in more detail later in this chapter.

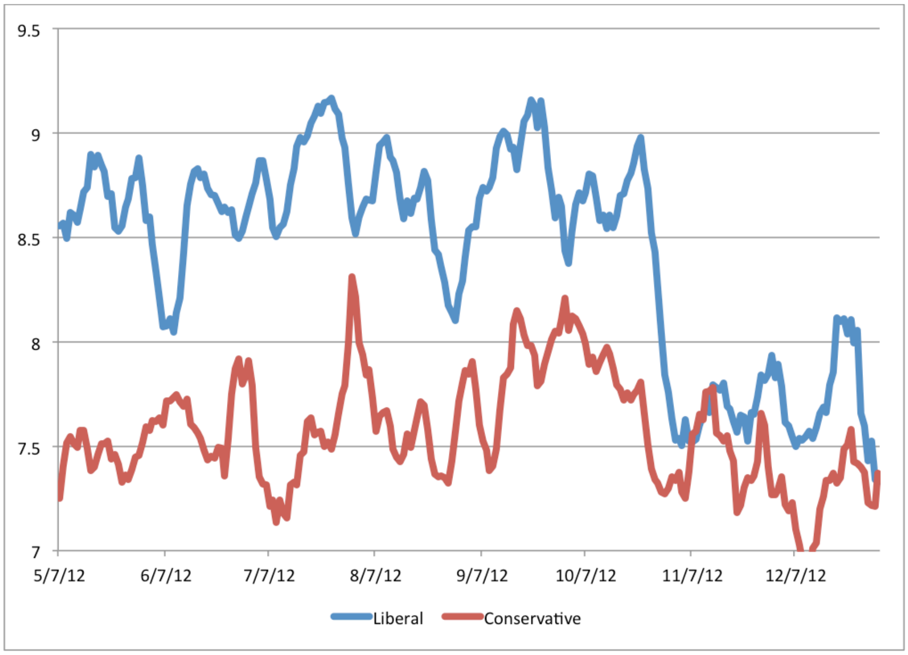


**Figure 6.1. Liberal and conservative use of pronouns over time.**

Figure 6.1 shows an example of a plot of liberal and conservative use of a linguistic style metric over time, in this case, pronouns (I, you, they, him, it), with the y axis being the percentage of words per text that are pronouns. Each data point is the average percentage frequency of pronouns for a given day, as written liberal or conservative blogs respectively. As we have seen in previous chapters, it is apparent that there is a difference of means between liberals and conservatives—in the above example, a paired t-test reveals a t-value of 30.314, with a p-value of 0.000, and the mean of the differences at 0.861.

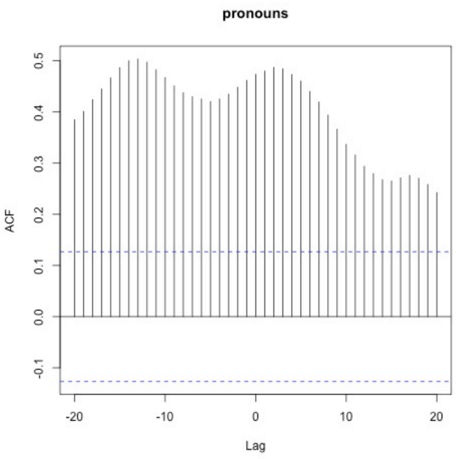
### *6.4.1 Function words*

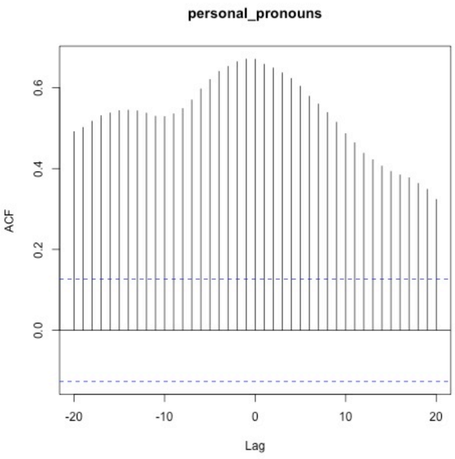
To address our central research question, we then tested whether liberal and conservative use of each of the LIWC measures was correlated over time by performing a cross-correlation analysis. For each measure, we took the seven-day moving average of the time series of function word use in liberal and conservative blogs.

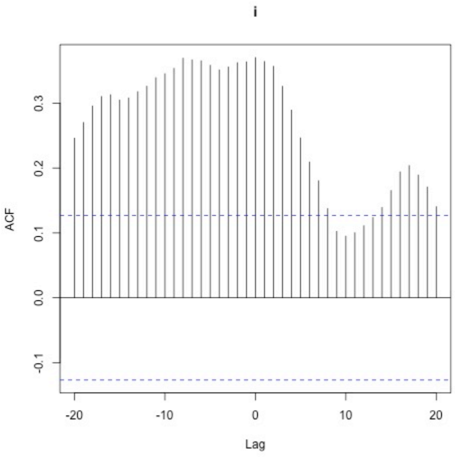
We found that function words are correlated positively between liberals and conservatives over time, and sometimes (for example, with pronouns) very positively. The results, shown in Table 6.2 below, were very consistent. Out of 21 possible categories of non-content words, at zero lag, 16 were positively correlated between liberal and conservative bloggers over time, four measures had no significant correlation, and one measure was negatively correlated.

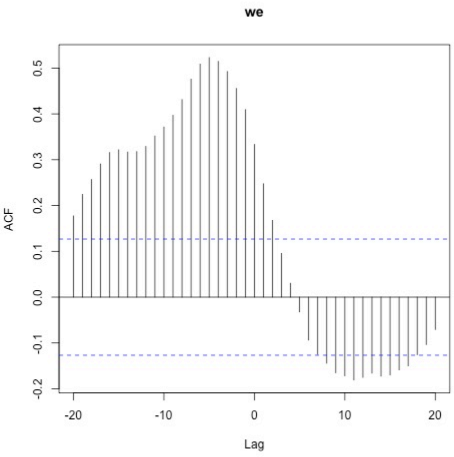
| Category | Correlation Coefficient (at lag=0) |
|---|---|
| Pronouns | 0.474** |
| Personal pronouns | 0.671** |
| I-words | 0.370** |
| We-words | 0.333** |
| You-words | 0.462** |
| She/he | 0.803** |
| They | 0.183* |
| Imp. pronouns | 0.242* |
| Articles | 0.316** |
| Verbs | 0.128 |
| Auxiliary verbs | -0.061 |
| Past tense | 0.693** |
| Present tense | 0.418** |
| Future tense | 0.294** |
| Adverbs | 0.333** |
| Prepositions | -0.114 |
| Conjunctions | 0.099 |
| Negations | 0.322** |
| Quantitative words | 0.245* |
| Numbers | 0.458** |
| Assent | -0.237* |

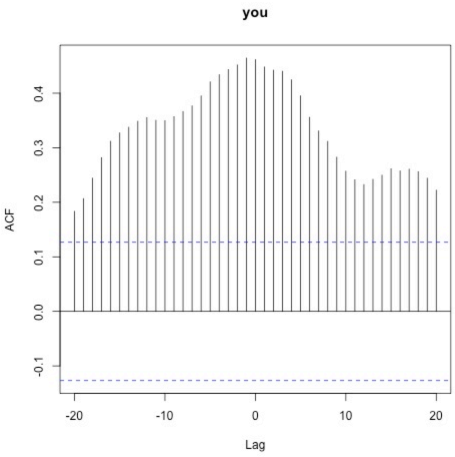
**Table 6.2**. **Table of correlations between liberal and conservative use of types of linguistic markers.**

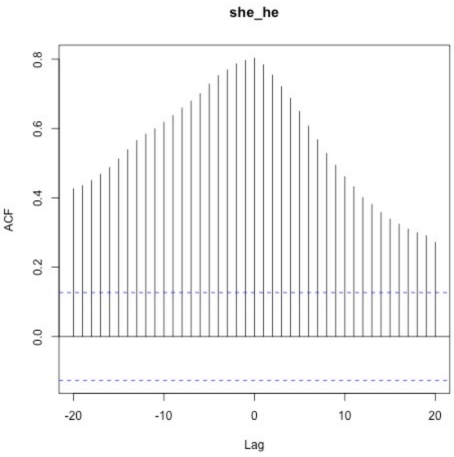

pronouns
ACF
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-20
-10
0
10
20
Lag

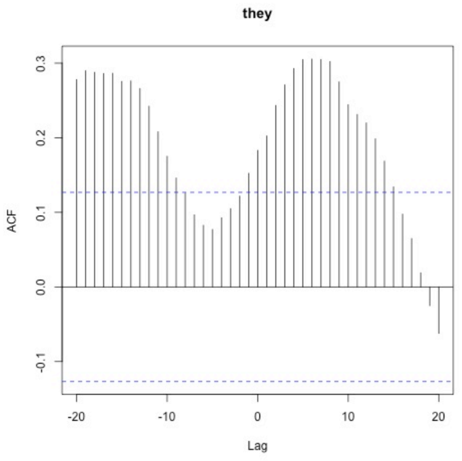

personal_pronouns
ACF
0.6
0.4
0.2
0.0
-20
-10
0
10
20
Lag

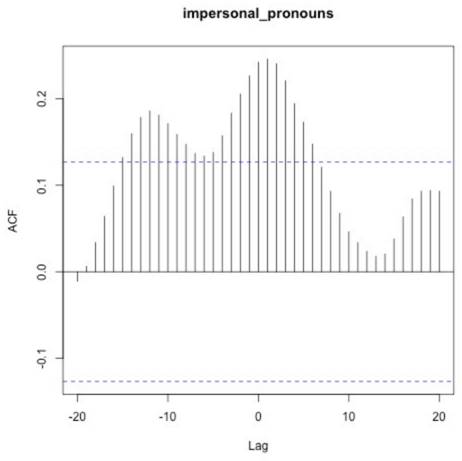

i
ACF
0.3
0.2
0.1
0.0
-0.1
-20
-10
0
10
20
Lag

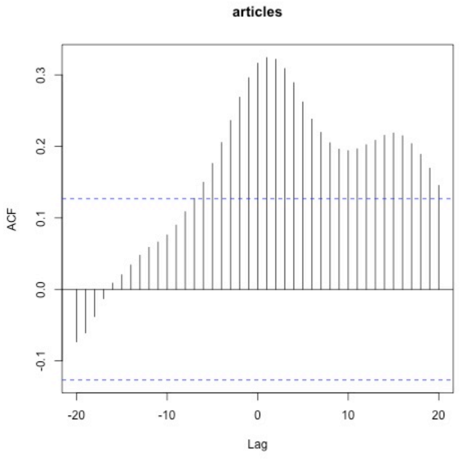

we
ACF
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-20
-10
0
10
20
Lag

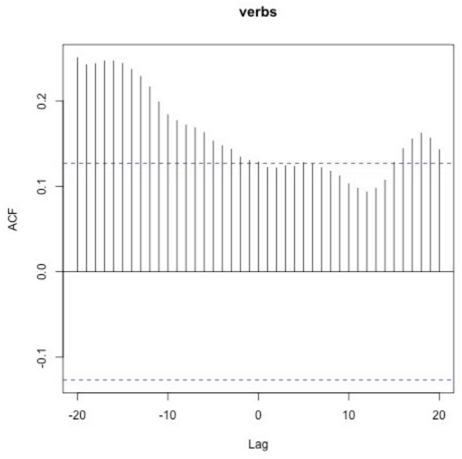

you
ACF
0.4
0.3
0.2
0.1
0.0
-0.1
-20
-10
0
10
20
Lag

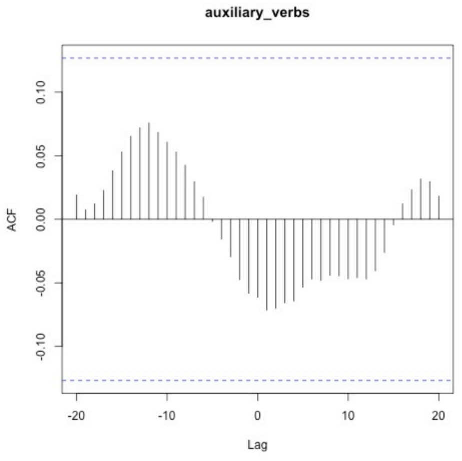

she_he
ACF
0.8
0.6
0.4
0.2
0.0
-20
-10
0
10
20
Lag

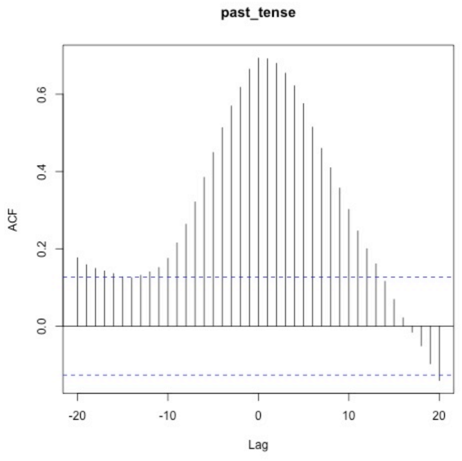

they

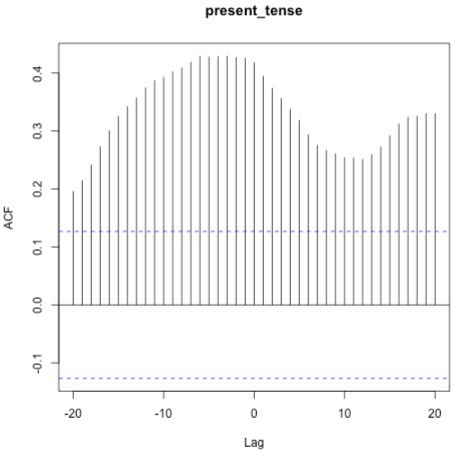

impersonal_pronouns

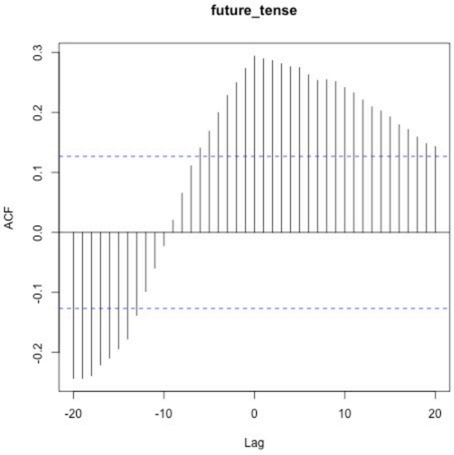

articles

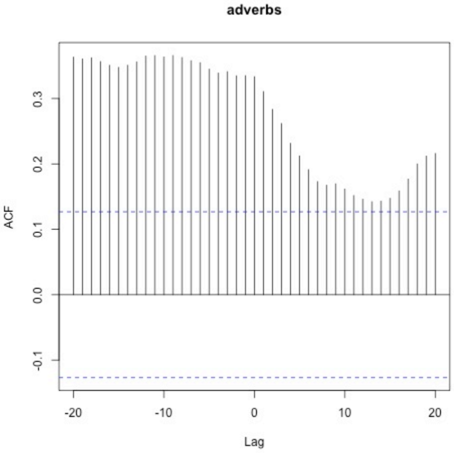

verbs

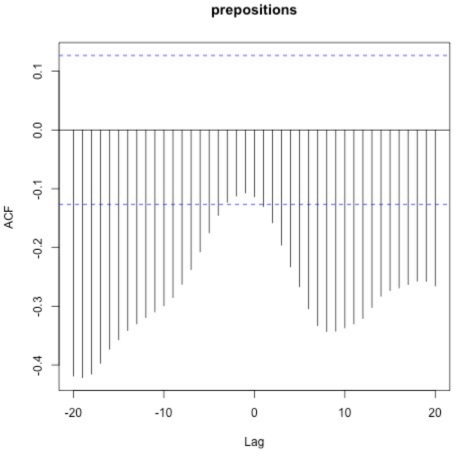

auxiliary_verbs

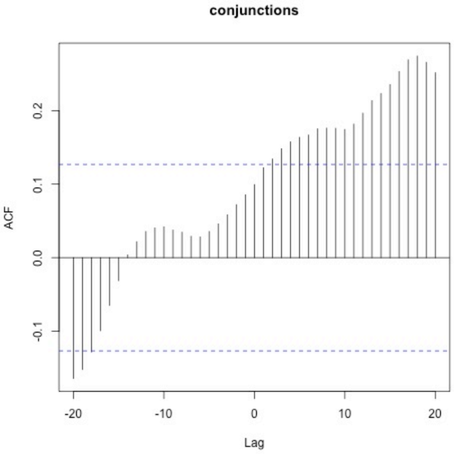

past_tense

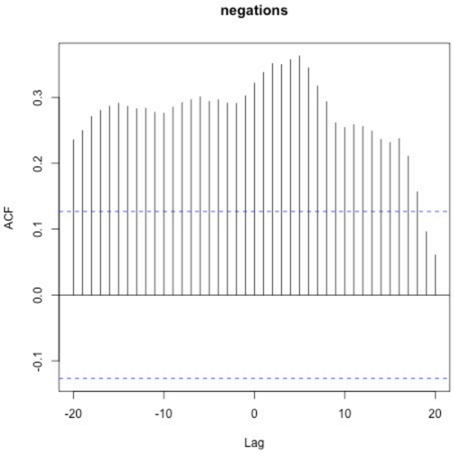

present_tense
ACF
0.4
0.3
0.2
0.1
0.0
-0.1
-20
-10
0
10
20
Lag


future_tense
ACF
0.3
0.2
0.1
0.0
-0.1
-0.2
-20
-10
0
10
20
Lag


adverbs
ACF
0.3
0.2
0.1
0.0
-0.1
-20
-10
0
10
20
Lag


prepositions
ACF
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-20
-10
0
10
20
Lag


conjunctions
ACF
0.2
0.1
0.0
-0.1
-20
-10
0
10
20
Lag


negations
ACF
0.3
0.2
0.1
0.0
-0.1
-20
-10
0
10
20
Lag

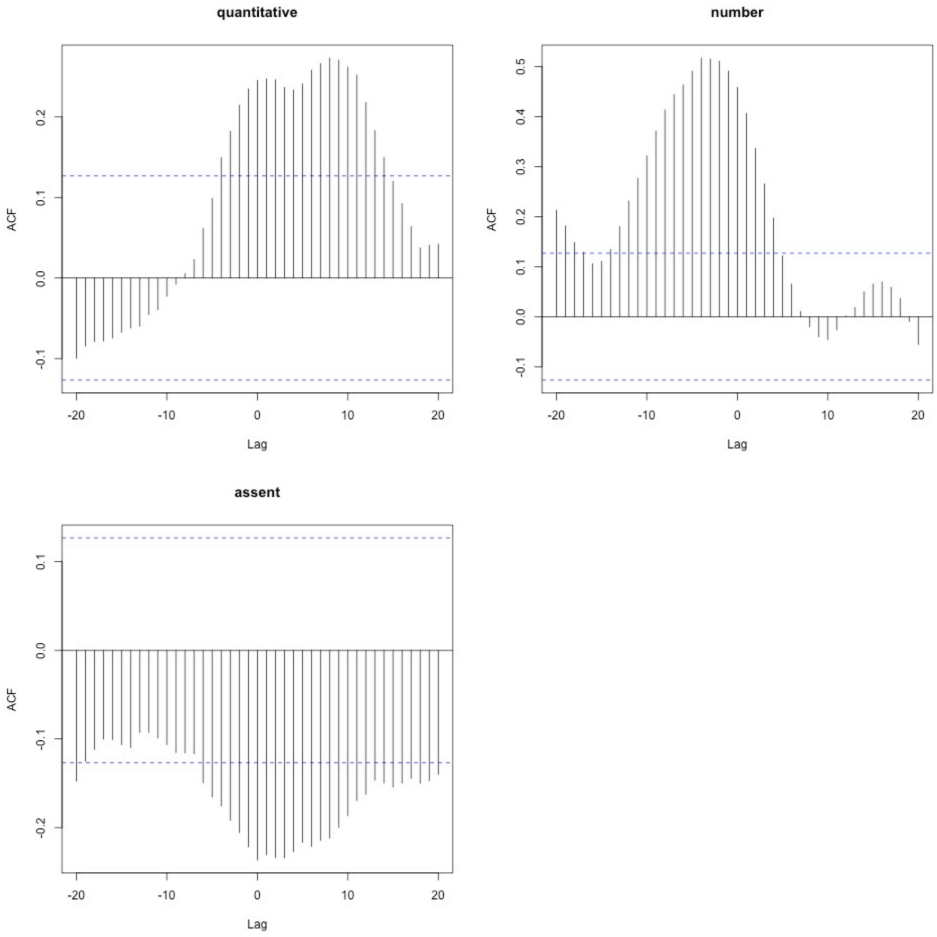


**Figure 6.2. Cross-correlation between liberal and conservative use of linguistic markers.**

Above are examples of some of the cross-correlation functions we found. The dotted line indicates the cutoff at which the correlation found is significant. Under the null hypothesis of no correlation, the variance of the cross-correlation function is about 1/*n* where *n* is the length of the series. As such, the critical value at the 5% level is calculated at $\pm 2/\sqrt{n}$, and in our case with *n* = 240 we determine that cross correlation coefficients over 0.129 are significant. It is further important to note that the measures in LIWC are normalized by word count, so verbosity of posts is likely not a confounding factor. To determine correlation as shown in  above, we report the cross-correlation coefficient at lag = 0, though it was interesting to see that there were often significant correlations at other lag values. Of the four measures for which we report no correlation at zero lag, three (verbs, conjunctions, and prepositions) had significant correlations at nonzero lag. In addition, though many of the above cross-correlation functions display maxima at or near lag = 0, the strength of the correlation between liberal and conservative presentation of some of these linguistic markers over time is sometimes much greater at nonzero lag (see the bimodal plot for "they" words).

Supposing that timing may have played a role in correlations in the frequency of usage of function words, we segmented our data set into various time periods to see if significant cross correlations during these subsets of time. The first segmentation we employed was between pre- and post- election blog posts, speculating that blogs may have been particularly influenced by the news media focus on the election, and that this may have influenced blogging behavior and therefore language production. We checked for cross-correlation in use of function words between liberal and conservative blog posts before the election and did the same for liberal and conservative blog posts after the election. This results of this analysis are shown in Table 6.3 below.

| **Category** | **Correlation Coeff. (lag = 0)** | | |
|---|---|---|---|
| | **Full range** | **Pre-election** | **Post-election** |
| Pronouns | 0.474** | 0.103 | -0.004 |
| Personal pronouns | 0.671** | 0.281* | 0.282* |
| I-words | 0.370** | -0.146 | 0.172 |
| We-words | 0.333** | 0.329** | -0.371** |
| You-words | 0.462** | 0.181* | 0.310** |
| She/he | 0.803** | 0.606** | 0.486** |

|  |  |  |  |
|---|---|---|---|
| They | 0.183* | 0.130 | 0.123 |
| Imp. pronouns | 0.242* | 0.152* | 0.167 |
| Articles | 0.316** | 0.346** | 0.699** |
| Verbs | 0.128 | -0.235* | -0.360* |
| Auxiliary verbs | -0.061 | -0.277* | -0.307* |
| Past tense | 0.693** | 0.532** | 0.452* |
| Present tense | 0.418** | 0.242* | -0.466* |
| Future tense | 0.294** | 0.509** | 0.317* |
| Adverbs | 0.333** | -0.099 | -0.050 |
| Prepositions | -0.114 | 0.328** | 0.213 |
| Conjunctions | 0.099 | 0.202* | 0.194 |
| Negations | 0.322** | 0.251* | 0.141 |
| Quantitative words | 0.245* | 0.432** | 0.187 |
| Numbers | 0.458** | 0.454** | -0.001 |
| Assent | -0.237* | 0.346** | -0.339* |

**Table 6.3. Table of correlations between liberal and conservative use of types of linguistic markers, with data split before and after the election.**

Pre-election cross-correlations were significant at > 0.160, and post-election cross-correlations were significant at > 0.252. As shown in the table above, 17 of 21 measures showed significant correlation between the liberal and conservative corpora across the full data collection period, and a somewhat different set of 17 of 21 measures showed significant correlation before the election, of which 15 were positive correlations and 2 were inverse correlations. Only 11 of 21 measures were significantly correlated after the election, of which five, or about half, were inversely correlated. This could be because blogs were particularly keyed into election stories prior to the election and less so afterwards, but the results could also be the result of statistical significance occurring with a larger number of data points. To test this, we split the data into two equal periods of four months each, shown below in Table 6.4.

| Category | Correlation Coeff. (lag = 0), sig. > 0.176 | | |
|---|---|---|---|
| | Full range | May-August | September-December |
| Pronouns | 0.474** | -0.074 | 0.759** |
| Personal pronouns | 0.671** | 0.047 | 0.805** |
| I-words | 0.370** | -0.001 | 0.389* |
| We-words | 0.333** | 0.202* | -0.003 |
| You-words | 0.462** | 0.144 | 0.626** |
| She/he | 0.803** | 0.500** | 0.910** |
| They | 0.183* | -0.133 | 0.408** |
| Imp. pronouns | 0.242* | 0.205* | 0.650** |
| Articles | 0.316** | 0.458** | 0.490** |
| Verbs | 0.128 | -0.361* | 0.315* |
| Auxiliary verbs | -0.061 | -0.320* | 0.105 |
| Past tense | 0.693** | 0.360* | 0.795** |
| Present tense | 0.418** | 0.049 | 0.493** |
| Future tense | 0.294** | 0.705** | 0.251* |
| Adverbs | 0.333** | -0.116 | 0.504** |
| Prepositions | -0.114 | 0.080 | 0.157 |
| Conjunctions | 0.099 | 0.218* | 0.305* |
| Negations | 0.322** | 0.130 | 0.577** |
| Quantitative words | 0.245* | 0.514** | 0.274* |
| Numbers | 0.458** | 0.178* | 0.535** |
| Assent | -0.237* | 0.255* | 0.156 |

**Table 6.4. Table of correlations between liberal and conservative use of types of linguistic markers, with data split into two equal parts**

As shown above in Table 6.4, 12 of 21 measures were correlated across the liberal and conservative corpora for blogs authored from May through August 2012, two of which were inversely correlated, whereas 17 of 21 measures were correlated across the liberal and conservative corpora for blogs authored from September through December, none of which were inversely correlated. As function word use is in general more highly correlated in the latter segment, we are led to believe that the number of data points examined alone is sufficient to explain variation in how closely liberal and conservative function word use track one another.

To see if there are trends in how closely correlated function word use is between liberal and conservative populations, we split the data into four equal two-month periods, again beginning in May and ending in December. The cross-correlations between liberal and conservative function word use can be found below in Table 6.5.

| **Category** | **Correlation Coeff. (lag = 0), sig. > 0.239** | | | |
|---|---|---|---|---|
| | **May-Jun** | **Jul-Aug** | **Sep-Oct** | **Nov-Dec** |
| Pronouns | -0.457* | 0.057 | 0.453* | 0.361* |
| Personal pronouns | -0.012 | 0.034 | 0.400* | 0.723** |
| I-words | 0.134 | -0.256* | -0.176 | 0.688** |
| We-words | 0.196 | 0.461* | -0.023 | 0.287* |
| You-words | -0.145 | 0.244 | 0.327* | 0.489* |
| She/he | 0.811** | 0.429* | 0.728** | 0.587** |
| They | -0.300* | -0.119 | 0.457* | 0.348* |
| Imp. pronouns | 0.092 | 0.279* | 0.539* | 0.512** |
| Articles | 0.440* | 0.463* | 0.494** | 0.737** |
| Verbs | -0.111 | -0.553** | 0.196 | -0.112 |
| Auxiliary verbs | -0.082 | -0.549** | 0.091 | -0.179 |
| Past tense | 0.282* | 0.485** | 0.474* | 0.893** |
| Present tense | 0.326* | -0.355* | 0.722** | 0.125 |
| Future tense | 0.858** | 0.158 | -0.280* | 0.281 |
| Adverbs | -0.125 | 0.190 | 0.095 | 0.020 |
| Prepositions | -0.124 | 0.158 | 0.670** | 0.184 |
| Conjunctions | -0.276 | 0.194 | 0.334* | -0.214 |
| Negations | 0.205 | -0.339* | 0.580** | 0.211 |
| Quantitative words | 0.026 | 0.582** | 0.399* | 0.333* |
| Numbers | 0.173 | 0.290* | 0.558** | -0.105 |
| Assent | 0.471* | 0.121 | 0.533** | 0.264* |

**Table 6.5. Table of correlations between liberal and conservative use of types of linguistic markers, with data split into four equal parts.**

In the first two months, 8 of 21 measures have significant correlation in use with two being negative correlations, July and August find 12 measures with a correlation, of which 5 were negative, 16 measures with correlation between September and October of which one was negative, and 12 from November to December with no negative correlations. Here, we find that how closely liberal and conservative function word use track each other does change over time, with function word use most closely tracking in September and October, the months immediately preceding the Presidential election.

To explain why this might be the case, we employ topic modeling, which is a form of natural language processing well suited to analyzing large corpuses of text documents. Topic modeling reveals trends in the content of documents and allows us to group documents by those trends without the necessity of manual coding.

We used Latent Dirichlet Allocation (LDA)-based topic modeling, which has become popular in large part due to its effectiveness in discovering understandable topics (Steyvers and Griffiths 2007). Its flexibility has permitted its use for a variety of types of text collections: user reviews (Titov and McDonald 2008), email (Dredze et. al. 2008), and weblogs (Yano and Smith 2010).

The method generates topics, collections of co-occurring words within documents that are found across multiple documents; these topics are produced without a priori knowledge about semantic content, though the method is predicated on the premise that the content of the corpus is semantically related to the groups of words generated (Blei et. al. 2003). Topic modeling identifies the distribution of topics across documents, outputting a result not unlike that produced by qualitative coding of documents. Each document can contain multiple topics. The automation afforded by topic modeling allows for analysis of larger data sets than would be feasible through hand coding.

In this deployment of topic modeling, each individual blog post is defined as a “document”, which is the operating unit of analysis. Our method of topic modeling generates eight words per topic, and we are able to specify the number of topics to be generated. After generating topics, we reviewed a sample of the documents in which each topic appeared in order to corroborate that the implied subject matter was present. Generating topics is an iterative process of expanding or limiting the number of topics until meaningful results are produced. Once a set of topics was generated, a coder examined the set of words in each topic, identifying which of the topics have an identifiable semantic meaning. Once such meaning was identified, a random sample of posts in which each meaningful topic appeared was produced, and by examining these, a coder was able to confirm that the posts identified as containing a topic by the topic modeler corroborated our assessment of the topic’s meaning. We chose to have 30 topics in our model.

Liberal Topics

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| [t1] | job | workers | school | labor | unemployment | education | million | public |
| [t2] | growth | economic | economy | rate | china | fed | market | recession |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| [t3] | bank | spain | government | greece | debt | euro | germany | european |
| [t4] | obama | states | percent | poll | carolina | numbers | virginia | polls |
| [t5] | tax | cuts | income | spending | budget | plan | cut | fiscal |
| [t6] | share | send | posted | email | friendly | iran | marijuana | retweet |
| [t7] | news | fox | give | media | report | story | department | read |
| [t8] | week | city | event | today | night | days | blog | live |
| [t9] | race | obama | web | speak | romney | palin | mitt | sarah |
| [t10] | policy | public | post | book | important | article | university | political |
| [t11] | family | home | children | help | kids | thing | book | young |
| [t12] | walker | wisconsin | campaign | county | scott | money | governor | milwaukee |
| [t13] | romney | obama | mitt | president | campaign | ryan | presidential | republican |
| [t14] | gay | marriage | religious | church | god | sex | christian | school |
| [t15] | think | thing | point | want | lot | actually | media | isn |
| [t16] | women | bill | house | legislation | abortion | act | senate | committee |
| [t17] | american | america | political | history | country | movement | power | nation |
| [t18] | thing | guy | want | maybe | think | look | find | night |
| [t19] | court | law | supreme | federal | states | case | legal | justice |
| [t20] | business | company | bank | money | companies | bain | million | financial |
| [t21] | ohio | vote | black | voting | election | voter | white | florida |
| [t22] | war | military | israel | muslim | president | iraq | security | attack |
| [t23] | video | fat | film | show | women | movie | love | note |
| [t24] | republican | president | democratic | democrat | gop | election | obama | president |
| [t25] | government | money | class | economy | need | america | want | rich |
| [t26] | rape | akin | republican | abortion | women | immigration | rep | todd |
| [t27] | energy | climate | oil | water | change | food | global | gas |
| [t28] | health | care | insurance | medicare | medicaid | plan | obamacare | program |
| [t29] | health | care | medical | study | quality | cost | patients | cancer |
| [t30] | gun | police | violence | school | shooting | nra | control | weapons |

## Conservative Topics

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| [t1] | ryan | paul | nation | biden | republican | national | democrat | joe |
| [t2] | show | internet | food | today | live | fil | chick | video |
| [t3] | place | vote | submitted | votes | room | council | obama | political |
| [t4] | school | black | women | education | students | children | college | university |
| [t5] | media | news | story | video | press | cnn | left | post |
| [t6] | law | court | government | states | federal | constitution | supreme | rights |
| [t7] | liberal | book | filed | player | fascism | post | read | link |
| [t8] | city | california | san | home | local | county | high | area |
| [t9] | thing | think | want | isn | point | need | maybe | bad |
| [t10] | obama | president | barack | bush | america | clinton | white | administration |
| [t11] | american | god | family | warren | america | history | love | father |
| [t12] | israel | muslim | egypt | israeli | hamas | arab | jewish | islamic |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| [t13] | gun | police | crime | shooting | control | violence | drug | law |
| [t14] | china | countries | united | foreign | country | south | chinese | russia |
| [t15] | attack | benghazi | security | libya | administration | ambassador | intelligence | rice |
| [t16] | political | change | fact | public | policy | thing | think | important |
| [t17] | military | war | iran | nuclear | defense | afghanistan | forces | weapons |
| [t18] | government | america | party | europe | political | economic | socialist | country |
| [t19] | republican | party | conservative | gop | tea | conservatives | democrat | akin |
| [t20] | house | republican | senate | democrat | bill | congress | president | reid |
| [t21] | health | care | government | obamacare | insurance | union | workers | pay |
| [t22] | marriage | gay | religious | left | speech | church | rights | anti |
| [t23] | tax | spending | debt | budget | government | cuts | income | federal |
| [t24] | campaign | county | money | council | group | million | political | convention |
| [t25] | job | percent | economy | economic | rate | unemployed | growth | report |
| [t26] | percent | election | obama | poll | voter | vote | democrat | romney |
| [t27] | energy | government | company | gas | oil | business | companies | green |
| [t28] | romney | obama | mitt | campaign | debate | 2012 | barack | posted |
| [t29] | department | report | general | information | holder | officials | attorney | house |
| [t30] | post | service | link | news | site | thanks | 2012 | mention |

**Table 6.6**. **Topics found in our liberal and conservative corpora.**

From this, we identified five liberal and five conservative topics that had to do with the election, in order to plot the prevalence of blogs that contained at least one “election” topic. The following topics were coded as having to do with the election:

### Liberal Election Topics

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| [t4] | obama | states | percent | poll | carolina | numbers | virginia | polls |
| [t9] | race | obama | web | speak | romney | palin | mitt | sarah |
| [t13] | romney | obama | mitt | president | campaign | ryan | presidential | republican |
| [t21] | ohio | vote | black | voting | election | voter | white | florida |
| [t24] | republican | party | obama | democrat | president | election | democratic | gop |

### Conservative Election Topics

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| [t1] | ryan | paul | nation | biden | republican | national | democrat | joe |
| [t3] | place | vote | submitted | votes | room | council | obama | political |
| [t26] | percent | election | obama | poll | voter | vote | democrat | romney |
| [t28] | romney | obama | mitt | campaign | debate | 2012 | barack | posted |
| [t24] | campaign | county | money | council | group | million | political | convention |

**Table 6.7**. **Topics found in our liberal and conservative corpora that pertain to the Presidential election.**

We then identified all blog posts containing at least one of these topics, making the assertion that those blogs were at least in part about the election. Using this code, we identified the daily proportion of blog posts containing at least one election topic and obtained the following plot:

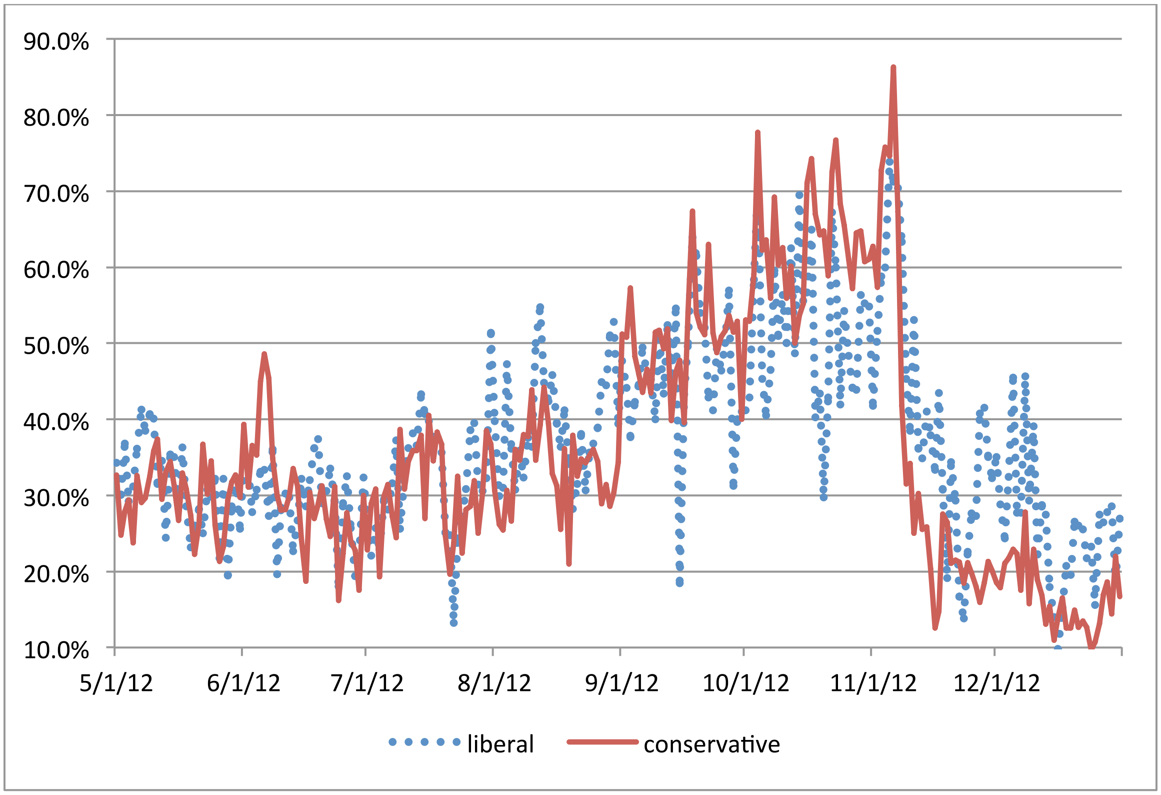


**Figure 6.3. Plot of prevalence of election topics within the liberal and conservative blogospheres.**

As shown in the plot above, in the first four months leading up to the election, about 30% of blog posts were about the election. As the election approached, and in particular during September and October, the liberal and conservative blogospheres began to converge on the election as a topic of discussion, with election topics on some days being contained in over 70% of all posts. Unsurprisingly, there is a dramatic dropoff in election coverage soon after November 6, the date of the general election.

While the generation of the groups of words referred to as topics in topic modeling is purely statistical and does not involve human input, the interpretation of these words is dependent upon human interpretation. It is important to note that it is highly unlikely that every single blog post coded under this method as pertaining to the election actually would be coded by a human as being about the election, and it is also highly unlikely that every single blog post that a human would code as pertaining to the election was identified as such by our topic modeling. However, the method is useful in showing us trends and giving us insight into our corpus that would not have otherwise been possible without an impracticable amount of manual work.

When we compare our election topic results with our cross-correlation results, particularly those results with two-month segments, we are led to surmise that a possible explanation for the common use of function words is the convergence of both the liberal and conservative blogospheres around election topics.

| | May-Jun | Jul-Aug | Sep-Oct | Nov-Dec | Nov. 7 – Dec. 31 |
|---|---|---|---|---|---|
| Lib. Election Topic Prevalence | 30.3% | 35.7% | 48.9% | 33.3% | 28.8% |
| Con. Election Topic Prevalence | 29.9% | 31.9% | 56.7% | 25.3% | 18.4% |
| Positive Cross-Correlations in Lib/Con Function Word Use | 6 / 21 | 7 / 21 | 15 / 21 | 12 / 21 | 6 / 21 |

**Table 6.8**. **Comparing election topic prevalence with liberal/conservative function word use correlation.**

Shown above in Table 6.8 is a comparison of the average prevalence of election topics in the liberal and conservative blogospheres in two-month groupings beginning in May. We compare these data with the positive cross-correlations we found in liberal and conservative function word use, also split in two month periods, and we find that during a period where discussions about the election are significantly more prevalent among liberal and conservative bloggers, there is far greater cross-correlation in liberal and conservative function word use. We note that a relatively high number of function words were found to be cross-correlated in November to December. We speculated that this could have been a product of the extremely high prevalence of election topics in the first week of the month of November, and we recall our earlier analysis that had excluded this first week. The results of the removal of the first week of November is dramatic—only six positive cross-correlations in function word use were found in the remainder of the year, as opposed to twelve when the first week of November is retained in that segment.

This finding suggests that a possible explanation for liberal and conservative blogs more similar use of function words stems from the convergence election topics we were able to observe.

### *6.4.2 Assent*

A small finding is that there is a significant negative correlation between liberal and conservative use of assent words. Assent words include "agree, accept, yes, OK." When liberals are using more words of agreement, conservatives are using fewer, and when conservatives are using more words of agreement, liberals are using fewer. The presence of a correlation is evidence that liberals and conservatives are reacting to a single stimulus as opposed to independently reacting to very different sets of stimuli. We speculate that the fact that the correlation is negative may be evidence that they may be framing these topics differently.

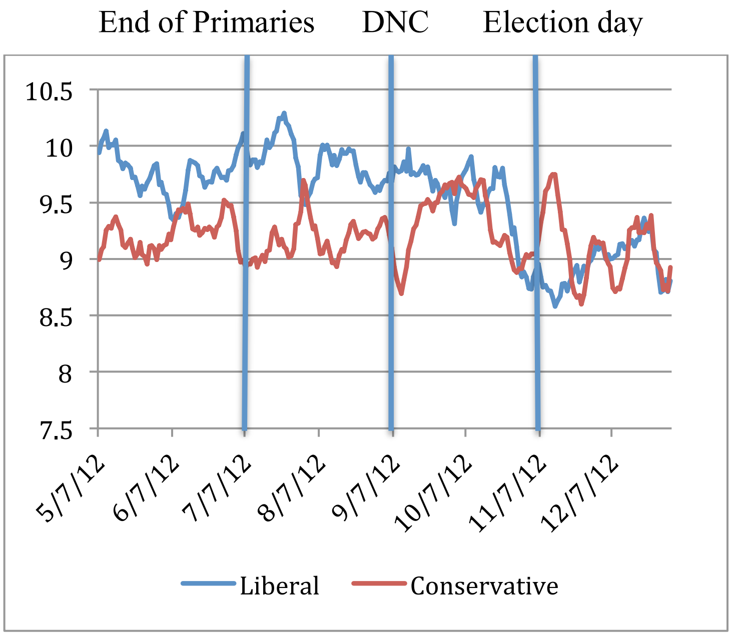


**Figure 6.4. Prevalence of assent-related words from liberal and conservative blogs.**

As seen in Figure 6.4 above, features on the graph of the prevalence of assent-related words align with real-world events. The Democratic National Convention took place in the days around the early September peak in liberal use of assent-related words. The next major peak of liberal use of assent words occurred around Barack Obama's victory. The Republican plot had fewer major features, but the flurry of assent-related words used by conservatives in early July correspond to Romney's official introduction as the presumptive Republican candidate following the official conclusion of the primaries.

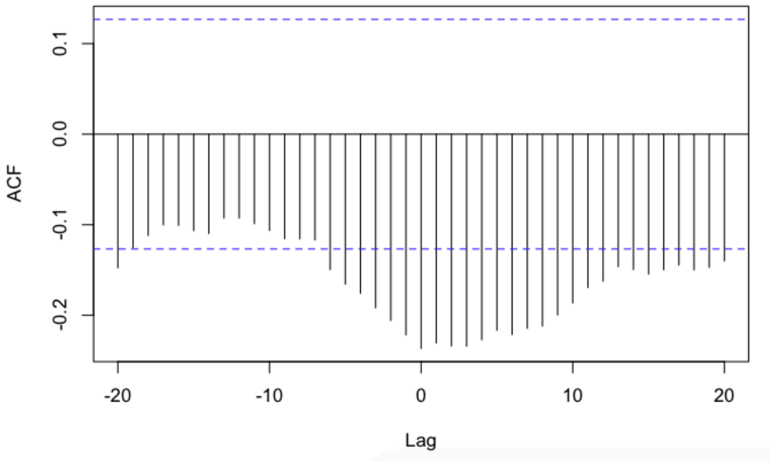


**Figure 6.5. Cross-correlation of conservative and liberal use of assent-related words.**

We speculated that this result could also be an artifact of the well-known "yes we can" slogan used by the Obama campaign which was developed during the 2008 campaign and which was still popularly used (and lampooned) during the 2012 election. As such, we performed a word count analysis on the phrase, finding that the phrase was used far less often by conservatives than by liberals. However, it may be the case that this liberal "rallying cry" was simply attested at different times from conservative appropriation of the phrase, leading to the negative correlation observed.

## 6.5 Discussion

In our analysis, we control for political ideology, separating liberals and conservatives into two separate populations for comparison. As we mentioned in the previous section, that there is a trend of positive correlation in liberal and conservative use of function words over the short term is a very interesting but difficult result—were we to find a *negative* correlation between liberal and conservative use of function words over time, we might be able to posit that political ideology mediated different reactions to environmental stimuli. However, function word use is *positively* correlated between these two groups over time. While the existence of a correlation in function word use within these two separate communities indicates that environmental and contextual factors are indeed influencing the use of function words, our primary manipulation—identifying posts as liberal or conservative in nature—could not explain the pattern we observed.

However, when we split the data into segments and examined the cross-correlation of function word use between liberals and conservatives over more limited time periods, it emerged that cross-correlations were more prevalent at different points in time, with a peak in the two months prior to the election. During this time, the two blogospheres were observed to have converged on the election as a topic of discussion, with a much higher percentage of blog posts dealing with the election.

While it is important to note that we cannot claim conclusively that the temporal differences we observed in the frequency of cross-correlations of function word use were the result of discussing the election, it is clear that the 2012 Presidential election was the defining political feature of the time period studied, and this is reflected in the prevalence of topics about the election, which peaked in the two-month period before the election. In the context of prior studies that show blogs and mainstream media have mutually deep influences on agenda-setting (Wallsten 2007), and studies of mainstream media that point to American election coverage historically reaching its peak around October and

November (Stone and Combs 1981), it is unlikely that this feature of the intensity of election coverage by blogs is a matter of happenstance.

Our finding is most interesting in contrast with the established understanding that function word use is largely independent of the subject being discussed, a claim made with the proposition that function word use is reflective of properties innate to the author (Pennebaker 2011). We think our results may add some nuance to this claim. We posit that when examined broadly and aggregated across time, function word use is likely related to those psychological properties identified as correlated with how people use the structure of language in the literature. However, our study shows that short-term fluctuations in function word use do exist and that these fluctuations are not purely random in nature. Instead, these short-term fluctuations may be driven by the subject matter at hand.

We recall Figure 5.1 from the last chapter, in which the syntactic representation of language is related not just to a speaker or author's situation model but also to the lexical representation—the content words that one chooses to use. This could mean that when speaking about the election, in addition to being more likely to say words such as "elections," "polls," and so forth, a speaker is also more likely to use conjunctions, prepositions, and pronouns in certain frequencies. We feel this to be an important corollary to the existing understanding of function word use, as the potential size of this effect is not yet understood. It could be that topic might be an effective tool for controlling variance when attempting to examine function word use in the future.

### *6.5.1 Smoothing*

We feel it necessary to comment on some considerations stemming from our methodological choices. As we noted earlier, we used a weighted moving average of the time series data for our analysis of cross-correlations. Data smoothing is a common practice in computational linguistics, as linguistic measures often carry a significant amount of noise, causing correlations that exist to be obscured. Also, the phenomena that produce change in linguistic measures might not take place on a timescale as granular as a day—for example, a news story on a certain topic might prompt blog posts on that topic over the course of a week. We noted that several others (cf O'Connor et. al. 2010, Gilbert and Karahalios 2010) have used moving averages in similar studies involving noisy linguistic data from social media.

However, it is important to take care when correlating data that has been smoothed, as the points within a smoothed data set are, on the whole, closer to the mean value of all points within the set. Comparing two data sets that have been altered in this way can create false correlations. Recognizing that perhaps the most obvious possible explanation for the significant correlations we found in function word use nearly across the board could be the confounding effects of data smoothing, we took a closer look.

Seeking to avoid drawing false conclusions based on falsely correlated data, we used a simple test—we took data from each measure and reordered them randomly such that any data point had an equal chance of being before or after any other data point. Afterwards, we applied smoothing to the newly randomized data series. We then checked the randomized-order data series against the result of the time-ordered series. An example of this is found in Figure 6.6 below.

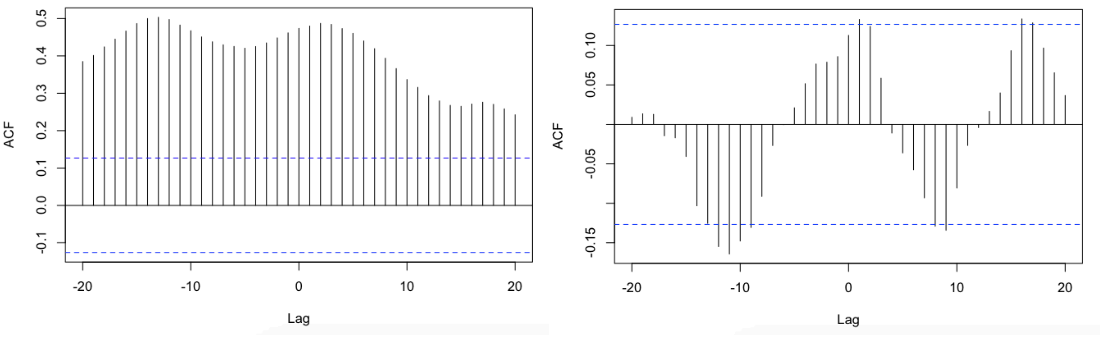


**Figure 6.6. Cross-correlations between liberal and conservative use of pronouns, weighted moving average. Left: original time series; Right: Order randomized.**

As seen in Figure 6.6, the time-ordered data for liberal and conservative use of pronouns was highly correlated, whereas the randomized-order liberal and conservative data were not significantly correlated. Note the different scales used. As we tested across the various measures, randomized-order data did not show significant cross-correlations. As such, we are comfortable rejecting temporal smoothing as the primary explanation for our results.

### *6.5.2 Interdependence*

We also considered the fact the various categories of non-content words are, to some extent, interdependent. Use of pronouns is syntactically associated with use of verbs, for example, and so perhaps it is less surprising that our results were so consistent across metrics. However, this does not explain why our style metrics were correlated across these two disparate populations.

### *6.5.3 Stylistic invariance*

As previously noted, few researchers have looked at short-term changes in linguistic markers, and those that did examined changes over time did not look at non-content words (cf O'Connor et. al. 2010) or did not use a time series analysis (cf Tausczik and Pennebaker 2010). Earlier, we noted that some researchers believe linguistic style to be relatively invariant. Based on our results, we believe that this assumption needs to be qualified to address both variance observed in the short term and in the long term. While it may be true that linguistic style is consistent within subjects in a long-term, aggregate manner, we observe here patterns of stylistic changes that appear to take place across entire populations and even crossing community boundaries. Based on the literature and our results from the previous two chapters, we believe that linguistic style is deeply influenced by psychological factors such as gender and ideology. However, we must add nuance to our understanding of linguistic style with the following:

- Linguistic style is *not* invariant on a short-term basis.
- Short-term fluctuations in linguistic style are necessarily not solely the product of individual, random contextual factors; rather, environmental stimuli that affect entire populations can have discernible effects on style even in the short term.

As research on linguistic style deepens and the breadth of its applications expands, addressing known contributors to variance will be critical. This is particularly so, as one of the major limitations of research on style is the "noise" observed in natural language corpora obtained from the wild. By better understanding the mechanisms that contribute to variance in style metrics, it is easy to imagine how textual analysis methodologies might improve—mediating factors that can be controlled for could very well be hiding in plain sight.

Stimuli that could be considered include psychological states such as stress, which has been shown to fluctuate over a time period of days (Horowitz et. al. 1979), or perhaps emotion. Fortunately, LIWC offers a way to gain some insight into the emotional language used by our two populations.

### *6.5.4 Affect*

To examine how liberals and conservatives bloggers expressed emotion, we used two measures provided by the LIWC dictionary to create a measure of sentiment—positive emotion words and negative emotion words. LIWC identifies words such as "happy," "pleasant," and "joyous" as being positive emotion words, and words such as "cried," and "abandon" as containing having negative emotions. We calculate a measure of polarity:

$$polarity = \frac{[p - n]}{p + n}$$

where *p* is the proportion of positive words in a document and *n* is the proportion of negative words in a document. A positive polarity indicates that there are more positive words than negative words, and a negative polarity indicates the opposite. The range is [-1,1].

Figure 6.7, below, depicts the weighted moving average of the average polarity displayed by liberal and conservative bloggers respectively.

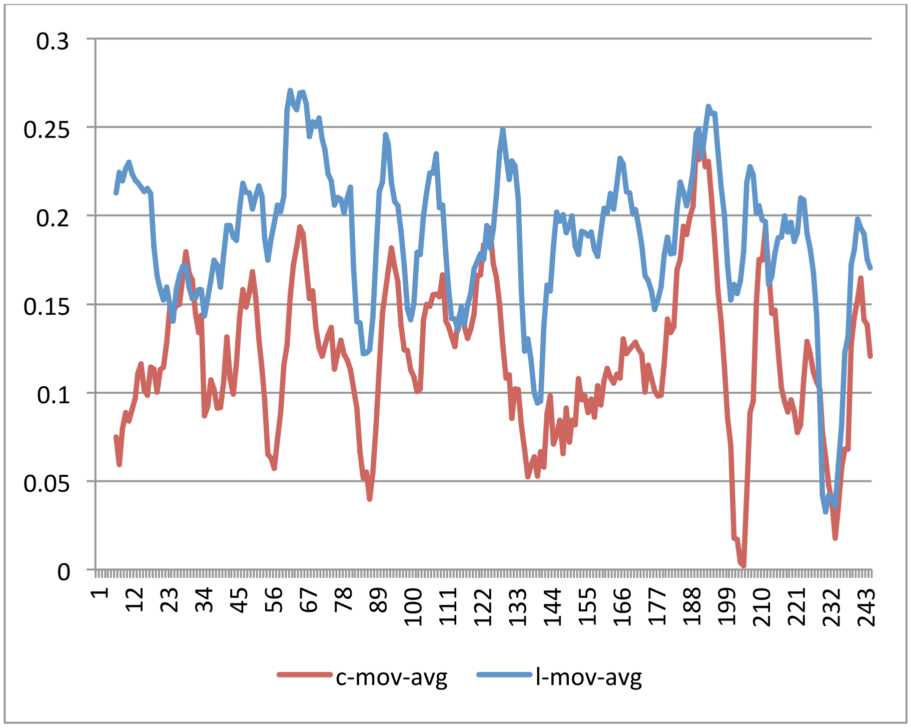


**Figure 6.7. Time series of polarity for liberal and conservative bloggers.**

We then calculate the cross-correlation function between liberal and conservative polarity as a function of time, shown in Figure 6.8 below. This graph shows a very significant positive correlation ($r$ = 0.519 at lag=0) between liberal and conservative polarity at any given time. In plain language, liberals and conservatives may be reacting emotionally to the same stimuli in the same way: conservatives demonstrate more happiness when liberals demonstrate more happiness, and vice versa.

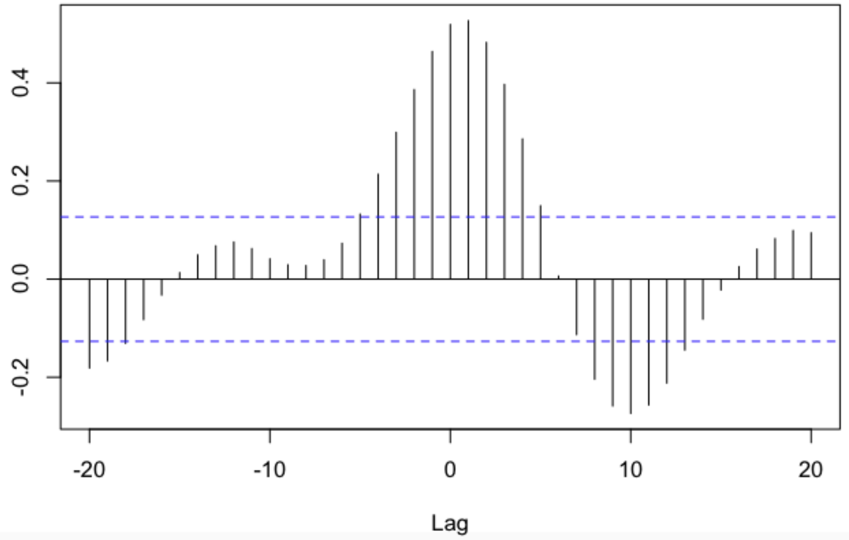


**Figure 6.8**. **Cross-correlation function of polarity for liberal and conservative bloggers**

However, other factors in the data could have led to these results. We examine the possibility of natural periodicity in affective language use below using periodograms of the time series for liberal and conservative polarity. High values in the periodogram indicate a strong periodic element at the given frequency.

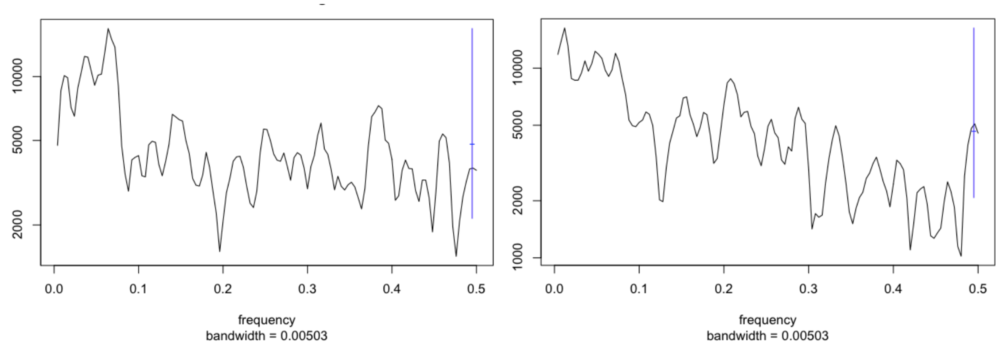


**Figure 6.9. Left: periodogram for conservative polarity; Right periodogram for liberal polarity**

The lack of strong local maxima in the above figures suggests that periodicity is not the key factor in producing the correlation found.

While these results on emotional language are interesting, we have failed to eliminate emotion as a possible mediating factor for function word use, and if it is a mediating factor, the presence of a causal relationship is far from confirmed. We are also left with perhaps more of a puzzle than we started with—liberals and conservatives are displaying positive and negative emotions in tandem with one another in addition to displaying stylistic changes in tandem.

## 6.6 Summary and Limitations

We can only infer from our results that some external or environmental factors affect the way in which function words and affective words were used by political bloggers as a whole. Based on the fact that we were able to detect correlations with a 7-day moving average, we may speculate that these factors operate within a scale of days. While we can not yet offer concrete explanations, our research is suggestive that systematic, experimental examination of stimuli that may affect linguistic style in the short term could very well bear fruit. Interpreted in the context of our prior findings on differences in liberal and conservative psychology, it is interesting that the prevalence of function words between these two groups had a positive correlation rather than a negative one, in light of the significant, time-invariant differences we found in the liberal and conservative use of these words in earlier chapters and the contrasting psychological underpinnings of the groups those differences suggest.

Probably the most obvious limitation to this study is that we are unable to definitively determine the etiology of the observed fluctuations over time of any of the LIWC measures, as the scope of our investigation limits us to speculation on that matter. In truth, we are not even certain of how we might even categorize the types of phenomena that might influence changes in the short term in something like pronoun use. We must also be cautious in making claims of validity beyond our population of political bloggers. It is likely that influences on bloggers are environmental and multifaceted, and their mechanism of action requires further study. We suspect that a simple explanation for the observed fluctuations may not exist, and that multiple sets of different factors might each provide partial explanations for our observations.

That said, we may consider the possibilities as we currently see them. Liberals and conservatives indeed sometimes write about similar issues contemporaneously, and individual topics may tend to lend themselves to certain types of narrative structure. Some stories, perhaps those about people, could lend themselves more to pronoun use, whereas others simply involve more preposition use. For example, as a story about the Federal Reserve becomes popular in news media, both liberals and conservatives blog about that story, which causes bloggers to invoke the phrase “despite the Fed’s inaction” more often in their writings. This and other similar phenomena could aggregate to form the correlations we observe. In this way, the prevalence of content words triggers variation in non-content words, and a directional causative relationship is indicated.

However, the idea that the topic of a text has an effect on function word use is one that has to date either been rejected, largely ignored, or assumed to be insignificant in previous studies on function words (Pennebaker 2011). If our finding lends credence to the possibility that topics or subjects have effects on function word use that are as significant as those we discovered, it is possible that earlier findings on function word prevalence may need to be reexamined for this confounding factor. To effectively test this, a reliable way of coding topics will need to be employed. It is unclear whether latent semantic analysis will be sufficient, or if manual coding of texts will be necessary, but either way, we believe the additional exploration of the relationship between content- and non-content words is a critical next step to our research.

While this may at first glance seem to be the simplest explanation, we must explore the implications of accepting it. First, liberal and conservative blogs have been demonstrated not to link to one another and be entirely separate communities. That they would write about the same topics at the same time—so noticeably that we detect significant correlations in function word use—tells us that blog content is either externally driven (blog topics tend to be those found in news), mutually driven (liberals and conservatives actually do tend to write and respond to the things the other side writes about), or both. Implications abound for our understanding of the relationship between blogs and news media, with blogs having been argued as driving topics in news media and vice versa (Wallsten 2007). This finding would lend support to the argument that there is a complex, bidirectional relationship in agenda-setting between blogs and newsmedia.

That liberals and conservatives might write about different topics and vary their writing style in the same way would be an interesting claim in itself as well, given the established understanding of differences in framing that liberal and conservative writers take to their stories. The samples we read of liberal and conservative writings on a given subject were strikingly different from one another, as can be expected, due to differences in framing. Framing differences consist of different moral, emotional, and normative stances taken in a text (Chong and Druckman 2007). Given this, our result would mean that the framing of stories—for example, news about Obamacare, a topic on which liberals and conservatives differed tremendously—alters the use of function words with an insignificant or much smaller effect size than that caused by the news topic itself.

An alternative explanation is that the content of posts is not the external factor causing the measures to vary with one another, and that something else is triggering the same responses in our liberal and conservative populations. If the variation in the frequency of non-content words is not an innate linguistic correlation between certain non-content words and certain content words as discussed above, this external factor must be triggering a psychological response in the authors of texts. We have already noted that psychological processes can drive changes in the prevalence of non-content words over longer time periods, and it is not absurd to suggest that they might be driving changes in the prevalence of these words over shorter time periods, as well. This is an intriguing possibility, but it is also maddeningly vague, as we cannot speculate as to the mechanism of these responses. We do know, though, that whatever the mechanism, the response of liberal and conservative bloggers to the stimulus (or stimuli) resulted in highly correlated changes across multiple non-content word measures. This suggests that despite the fundamental differences we established in Chapters 4 and 5 in liberal and conservative use of non-content words, the mechanism causing short-term fluctuations seems to be the same across both populations.

Further investigation is necessary to determine conclusively the cause of the fluctuations we identified. While we believe our results to paint a compelling picture that the prevalence of certain topics (the Presidential election in particular) may have had an effect on writers, it is difficult to establish a causal relationship when passively observing behavior as we have done. However, our central finding remains certain: that two distinct populations of bloggers reacted to extraneous events such that the structure of language used by these populations changed in similar ways, despite the significant endemic differences in these populations described in previous chapters.

Our results suggest that the findings of significant differences we found in the prior chapters of this dissertation need to be tempered by the realization that there is still much to be learned about what these function words mean.

# Chapter 7 : Discussion and Conclusions

## 7.1 Contributions

Naturally, an approach that is dependent on quantitative computational methods for data analysis has many limitations, among which are included a difficulty in demonstrating causative relationships between metrics for which there is correlation, and the possibility of confounding factors that may remain unknown. As such, this research should be positioned in two ways—first, as an effort to gather evidence in the field of politics, in which conflicts between pertinent theories have not yet been settled—and second, as an attempt to help broaden the social scientist's toolkit. It is our hope that despite the clear limitations of this study, we have contributed to extending the use of a promising computational tool that is still relatively new to the social sciences by having employed it in novel ways on a relevant, contemporary subject matter. We will discuss these contributions below.

### *7.1.1 Function Words*

The idea that the use of function words reflects some configuration of values or innate personality traits is an intriguing one—this "window to the soul," as it is referred to by Pennebaker, may become a critical tool in the data analyst's toolkit in years to come. In our research, we have found evidence that use of these words is indelibly tied to ideology and gender—deep psychological constructs at the core of our being. We, like others before us, believe in the power of these oft-ignored "junk" words to help us approach the increasingly available large data sets gathered from social media.

In this dissertation, we have found through examination of function words that liberals speak like more like women and conservatives speak more like men. We have found that the liberal and conservative blogospheres have coalesced around different styles that may reflect underlying psychological differences between people of different ideologies. Finally, we have found that liberal and conservative use of function words experiences a statistically significant covariation on a short-term basis, suggesting the existence of some external, environmental stimulus to which the production of language by liberals and conservatives reacts similarly—all while that same production of language reveals the deepest of differences between them. Clearly, function words are not without their complications.

In order that the interpretation of data regarding function words not be thrown further into confusion, in addition to exploratory research at the intersection of big data and textual analysis, other research approaches are necessary to supplement our understanding of function words. We think an experimental approach might be well-suited to this investigation, since part of the difficulties we encountered with analyzing data from the wild is the difficulty it poses in controlling for environmental and contextual factors. Does the use of function words vary when subjects are primed for certain psychological states? As no convincing taxonomy yet exists as to which psychological frames or other personal idiosyncrasies might have effects on an individual's use of function words, we hope exploratory work like ours may point experimentalists in fruitful directions for research.

### *7.1.2 On Politics*

We have sought to contribute to the ongoing discussion on the nature of political partisanship in the United States. In particular, we hold that our results provide further validation for moral foundations theory, which argues that ideology is not derived from a sum of various utility-maximizing political positions, but rather that political positions are informed by an ideology that is rooted in the core values that one holds.

One of the dominant social constructs that governs the way we live is gender—it is inculcated and constantly reinforced by parents, peers, the media, and society at large. In this dissertation we posit that the influence of gender extends into ideology, with our finding that conservatives write more like men and liberals write more like women. Our research calls attention to the surprising similarities between gender and ideology, and political positions—what is gender but an effort to create a normative personal reality, as directly informed by a series of prescribed values? What are political positions but efforts to create a normative societal reality, as directly informed by a set of prescribed values?

As ideological divides seem to deepen in the United States, it is all the more important that we seek to understand how and why these divides occur and what their implications may be. If it is the case, for example, that differences in linguistic style can lead to difficulties in interpersonal interactions, it is undoubtedly beneficial for this to be recognized from a scientific standpoint. First, popularizing the understanding that the way we have been trained to speak may be dividing us as a people is the first step to helping individuals be empowered to recognize and overcome their differences with one another. We hope that our research can add to a scientific narrative that urges the recognition of language as simultaneously extremely powerful and that our words, even as they have enormous impacts on how we are perceived, are shaped by forces of which we may not even be aware or are simply out of our direct control.

## 7.2 Methodological Reflections

Part of our contribution comes from the lessons learned about the quantitative methods we used, both theoretical and technical. The road to unlocking the promise of textual analysis that is backed by solid psycholinguistic theory may not be a smooth one—the difficulties we encountered throughout this data collection and cleaning process are revelatory of a problem that is faced by quantitative social science researchers who habitually deal with data obtained from the Internet—issues that fall under the categories of conventions and standards, contextuality, and glitches. We believe that textual data is inherently messy for a variety of reasons. To some extent, awareness of potential confounds can allow researchers to control for some of these issues, while other issues are not so easily addressed.

It is also appropriate to address the limitations of the tools we used as well. While numerous researchers have utilized LIWC to great effect, it is important to maintain a healthy skepticism towards the validity of its measures, especially given the specific context in which they are employed. The other obvious limitation of LIWC is that it only counts and classifies words. As we noted in a previous chapter, natural language is tremendously complex, and there is tremendous meaning encoded in language at the phrasal level, at the sentence level, topically, in subtext, and so forth. While some computational methods have begun to access these higher-level approaches to understanding language, we hold that particularly with regards to language, theory should seek consistency with a grounded, qualitative understanding of reality.

In our study, we examined the writings of liberals and conservatives for comparative purposes. In an experimental setting, we might have two groups, one liberal and one conservative each writing to a single prompt, after which we would compare the textual output of both groups. In analyzing blog posts in the wild, many more factors could potentially be at play. Were all political blogs writing only on politics? Is the presentation of a liberal / conservative dichotomy valid, or is political orientation on a continuum and multidimensional? How much does the fact that blogs quote one another and cross-post impact our findings? Though we were able to collect and analyze a large number of blog posts that served as a slice of the American political blogosphere, exploratory findings are ultimately most interesting when they inform more exhaustive, in-depth research.

While many studies of the structure of language have posited links between external conditions, psychological states, and use of functional words, studies to date have generally only compared changes in functional word prevalence across a small number of discrete time periods usually greater than a year. These prior studies often examined differences in language structure around an individual event, or between arbitrary time periods. In chapter 6, we compared observable, aggregate trends in the use of language on a much more granular timescale. We find that liberals and conservatives actually change their use of language in very similar patterns over time, implying the existence of some underlying environmental or contextual factors that affect our use of language on a short-term basis.

In an area of study that has not yet reached maturity, it is appropriate that our results raise more questions than they provide answers. Though we are unable to definitively couch results in an overarching theoretical framework, we have both confirmed and challenged aspects of the existing body of knowledge on computational linguistics and political partisan psychology. The importance of this research reaches beyond these disciplinary boundaries, however—the angle we took provides a window into what may become possible in analytics as we expand our understanding of psychology and the rich world of textual analysis.

### *7.2.1 Technical challenges for scraping*

The first categories of issues we faced came from interacting with the various types of formats and conventions that are involved in the delivery of textual data, such as text encoding standards, website structures, and date and time formats. We encountered significant heterogeneity in the way information was presented even within the singular medium of the political blog. Such issues have long since been identified in digital social research, in which researchers have been confronting the irony of vast arrays of impeccably ordered data whose ordering is unhelpful for or even impedes the possibility of its analysis (Marres and Weltevrede 2012).

This challenge to research emerges because public-facing renditions often are not formatted to be easily interpretable by computers—the visual and spatial cues that humans use to determine the context of the data that is being presented to them are, generally speaking, not available to scrapers, which use structural markers in website code to determine patterns of content. Oftentimes, these structural markers are intentionally obfuscated to enable legibility by humans and not by automated processes. One way to circumvent the difficulties is simply to access the desired data directly from databases, which is only possible if one has ownership, or if one is granted some portion of the privileges of ownership through the use of APIs, as we employed with the blogspot blogs. Ownership of data greatly simplifies the task of curating data and its accompanying metadata. In the absence of ownership of data, scraping serves to accomplish what direct database access otherwise would enable more easily.

The idea that gathering and analyzing textual data for quantitative analysis is “messy” or “dirty” is not new, with many having noted the onerousness of online data collection (Savage and Burrows 2007, Bollier 2010). The Internet contains a constantly evolving jumble of textual formats and Web technologies that make data collection and interpretation difficult. Errors we encountered as a result of encountering new standards had to be accounted for individually, a task involving human intervention that increases in difficulty as scope of analysis broadens.

### *7.2.2 Considerations of contextuality*

A deeply contextual but nevertheless important part of obtaining textual data from the field is deciding what is data and what is noise. This is a process that should take place deliberately, as false positives and systematic errors are easy to make due to the complexity of the data being examined.

We spent much time and energy checking the data for possible technical or contextual errors—articles that were blank, articles that were only excerpts from news articles. It can be difficult to automate the process of determining what data should be analyzed and what should not. Many researchers of natural language texts encounter similar issues: Back et. al. (2011), for example, attempted to use LIWC to identify anger-related messages sent to pagers. However, when they checked the LIWC data against their hand-coded ratings of anger, they found significant discrepancies. The reason for the discrepancies was found to be the transmission of system messages such as "critical error," which were parsed through LIWC's dictionaries and generated false positives for the anger measure. These system messages were so prominently featured in the data that it altered the correlations the authors were testing for. They suggest that "extremely careful" control routines are still required for the analysis of large digital textual data sets.

Our experience corroborates their findings: even if user-generated content can be distinguished from its backdrop, determining which content is useful is absolutely essential—for example, separating posts from comments is obviously important, since posts and comments have different authors (we were careful to do so in our analyses, discarding comments). Even if comments are the data one is interested in, this can cause difficulties of comments are being obtained from more than one blog: they are not always structured within a page in the same way from site to site.

These observations critically call our attention to an important aspect of digital social research—that *all parameters for collecting and formatting data are defined by the analyses that are to be performed*. We must consciously, at great effort and possibly expense, take already-formatted data and re-format it for our particular purposes. This act of curation deserves further examination because within our analytical tools lie assumptions about what is important and what is not. The discovery of the importance of function words already demonstrates the idea that something commonly ignored can actually be quite important and informationally rich.

The methodological issues that we face seem to originate from the deeply contextual nature of textual analysis. While other research may be interested in blog posts that have only video, in our case, video-only posts added essentially blank documents to our data that had to be removed so as not to influence the results. The use of digital media is messy—as many in CSCW have observed (cf Suchman 1995), people will inevitably find many unanticipated uses and styles of uses given any technological medium. As we seek to make the use of quantitative linguistic methods more of a “science of the artificial” (Ackerman 2000), we are reminded of something akin to CSCW’s socio-technical gap: while we are aware of many of the social factors that need to be addressed in the mechanical processes of data collection and interpretation, we find that for many kinds of studies some of these contextual considerations are not accounted for in the technical mechanisms for data collection and analysis and that significant manual intervention is required.

Certainly, the lessons that we learned are relevant to researchers studying textual data in the digital sphere, and the issues we identified, which include data ownership and usability, are also extremely relevant to those appropriating digital data for commercial purposes. In fact, it is likely that the technical and methodological challenges faced in digital social science research parallel those confronted by many non-academic actors in the digital sphere.

While it seems to be a foregone conclusion that companies providing online services will increasingly use user data available to them to accomplish any number of their goals, including marketing or improving their user experience, there have not been widespread attempts to commercialize use of data that is "owned" by others—that is, easily accessible, stored and structured in a way such that manipulation and analysis is allowed.

As the current trend is for data sets in digital social science research to get bigger—and since to enable this, a significant investment in accounting for the myriad conflicting technological standards is necessary, we call for the development of tools for data collection geared towards the needs of analytics and social science. A huge barrier to better understanding human behavior online and to using our understanding to create new technologies is the ease of availability of data for analysis. Better standardization, accessible APIs, better and more powerful scraping tools, better workflows for handling analysis of large data—all of the above will go a long way towards enabling the next generation of quantitative social science research online.

## 7.3 Future Directions

Part of what makes our line of research particularly compelling is the possibility that advances in computerized methods in natural language processing may be used to better understand and support users of online technologies. Already, data analytics of usage patterns from online applications have been implemented in ways that have great commercial and/or cultural value. NLP is already being used for consumer classification in targeted marketing (cf Kaefer et. al. 2005). It is a distinct possibility that more sophisticated approaches to applying NLP for commercial approaches will have tremendous economic and even societal impacts, as the latent information hidden in the wealth of publicly available textual data is unlocked and interpreted.

It would be interesting to consider the possibility that one might be able to monitor the textual output of many subjects in an automated way, testing it for linguistic style and other simple linguistic measures. We expect that as our understanding of computational linguistics and natural language processing broadens, these discoveries will be able to be applied in the workplace, by the government, and in commercial settings. The Linguistic Style Matching metric, for example, has been shown to be correlated with the likelihood of successful interpersonal collaborations. Our finding that political affiliation is significantly correlated with use of functional words and our finding that short-term environmental influences have effects on use of function words are suggestive of the existence of underlying psychological mechanisms that have to date been inadequately explored. Currently, few empirical studies at this time on which to base a convincing psycholinguistic model for these phenomena.

### *7.3.1 Quantitative Linguistics Research*

More empirical evidence must be amassed to confirm or debunk the results of researchers to date regarding the use of function words. As we've noted previously, function word use is associated with a wide array of cross-cutting demographic and social attributes, of which political orientation and gender are only two. Merely establishing that these categories have effects on function word use is not enough: the ultimate goal of scientific research is to explain and predict behavior. Of course, the first step to doing so is to accurately observe behavior and interpret it in the context of existing understanding, which we have attempted to do in a bounded way over the course of this dissertation.

However, for a universe as vast and complex as that of human language, many more measurements will be necessary to form a complete picture. We can only begin to speculate as to the rich, complicated, and chaotic psychological mechanisms behind a person's tendency to use (or not use) a few unobtrusive words. The movement of the planets was not computed based on one or two snapshots of the sky—it was only based on Brahe's efforts to make detailed measurements of the position of astral bodies that Kepler was able to formulate in his laws of planetary motion.

Natural language processing is still in its infancy, both in its methods and in our understanding of its results. Human language is vastly more complex than planetary motion, and there remain myriad applications and unexplored research questions for which the tools we used have not yet been applied, both in the study of politics and beyond.

The following list explores some research questions that are directly related to this dissertation:

1) It would be very valuable to further confirm the validity of our findings. Can the results of this dissertation be replicated in other Anglophone countries? Non-Anglophone countries? Can we observe the same phenomena in other social media, such as Twitter and Facebook? Are the results we found consistent over time? Do the findings change if we observe the 2008 election? 2016?
2) We would like to better understand the connections between linguistic markers and real-world behavior, particularly in the political sphere. Can online writings be used as a stand-in or supplement for polling? Are there linguistic markers that point to the inclination of voters to vote in one way or another?
3) We would further like to explore the tools that enable quantitative linguistic research on internet data. Is it possible, useful, and meaningful to create tools that

analyze word count measures on a real-time basis? What might we be able to learn by having access to word counts from social media in real time?

### *7.3.2 Research on Political Psychology*

Understanding the psychology of political partisanship involves understanding part of what makes us human—our adherence to our most deeply held beliefs—why we hold them and what we do with them. Our results support the thesis that political partisanship is more than just a personal utility-optimizing stance. Future research on partisanship that involves computational linguistics should involve exploring the possibility that analysis of writings can be revelatory of additional differences and similarities between people of different partisan orientations. Research questions might include:

1) The result of liberal and conservative writing having gendered characteristics should be further examined. Is Lakoff's gendered model of political psychology applicable to the politics of other countries? Is the strict father / nurturant parent dichotomy observably present in other countries?
2) We would like to further explore the idea of linguistic convergence in the political sphere. A possible experimental study could determine whether language that has the linguistic characteristics of text produced by liberals or conservatives is more appealing to one group or the other, even if it does not contain any content that reveals a political stance. In addition, the mechanism of linguistic convergence could be examined by observing the changes in an individual's textual output over time as they join and participate in communities such as the political blogosphere.

3) What is the real-world impact of linguistic differences between liberals and conservatives? Will liberals and conservatives conversing in a lab setting accommodate their linguistic style? Can they be primed to do so? Primed not to do so?
4) How can we capture differences in framing of issues by people of different political orientations? Does the use of affective language correspond to positive or negative views on a given issue? Can we use NLP to detect or otherwise observe political framing?

## 7.4 Conclusion

Our study supports the notion that there are critical differences in the way liberals and conservatives in America understand the world around them. We believe that in the field of computational linguistics, we have but scratched the surface. In our attempt to bridge the tools and understanding brought by relatively recent advances in computational linguistics with the extraordinarily complex body of knowledge on ideology and partisanship in sociology and political science, a few realizations are apparent: first, attempting to bind together the various traditions of knowledge within social science is tremendously valuable. The goal of all social science is, in the broadest sense, to understand and explain the behavior and structure of the human experience. Though establishing valid comparisons and connections across fields within social science is a very difficult task, unquestioned assumptions or assertions can be appropriately challenged or confirmed.

An important opportunity to advance the social sciences comes with the adoption of increasingly sophisticated statistical and computational tools. For example, unsupervised machine learning techniques such as latent semantic analysis hold particular promise for finding patterns in large patterns of text. Designing and properly applying and interpreting machine learning algorithms to social science problems requires a certain breadth and depth of understanding across multiple fields, a requirement that is a challenge to the dominant disciplinary boundaries in academia.

# Bibliography

Ackerman, M. S. (2000). The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human–Computer Interaction*, *15*(2-3), 179-203.

Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery* (pp. 36-43). ACM.

Alford, J. R., Funk, C. L., & Hibbing, J. R. (2005). Are political orientations genetically transmitted?. *American political science review*, *99*(02), 153-167.

Almquist, Z. W., & Butts, C. T. (2013). Dynamic network logistic regression: A logistic choice analysis of inter-and intra-group blog citation dynamics in the 2004 US presidential election. *Political Analysis*, *21*(4), 430-448.

Alpers, G. W., Winzelberg, A. J., Classen, C., Roberts, H., Dev, P., Koopman, C., & Taylor, C. B. (2005). Evaluation of computerized text analysis in an Internet breast cancer support group. *Computers in Human Behavior*, *21*(2), 361-376.

Back, M. D., Küfner, A. C., & Egloff, B. (2011). "Automatic or the people?" Anger on September 11, 2001, and lessons learned for the analysis of large digital data sets. *Psychological Science*, *22*(6), 837-8308.

Baldassarri, D., & Gelman, A. (2008). Partisans without constraint: Political polarization and trends in American public opinion. *AJS; American journal of sociology*, *114*(2), 408.

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, *18*(2), 135-160.

Barcus, F. E. (1961). A content analysis of trends in Sunday comics, 1900-1959. *Journalism & Mass Communication Quarterly*, *38*(2), 171-180.

Bauer, M. W., & Gaskell, G. (Eds.). (2000). *Qualitative researching with text, image and sound: A practical handbook for social research*. Sage.

Baum, M. A., & Groeling, T. (2008). New media and the polarization of American political discourse. *Political Communication*, *25*(4), 345-365.

Beer, J., Jeffrey, N. O., & Frohnen, B. (2014). *American Conservatism: An Encyclopedia*. Open Road Media.

Begel, A., DeLine, R., & Zimmermann, T. (2010, November). Social media for software engineering. In *Proceedings of the FSE/SDP workshop on Future of software engineering research* (pp. 33-38). ACM.

Bell, A. (1984). Language style as audience design. *Language in society*, *13*(02), 145-204.

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. Yale University Press.

Benoit, W. L., Hansen, G. J., & Verser, R. M. (2003). A meta-analysis of the effects of viewing US presidential debates. *Communication Monographs*, *70*(4), 335-350.

Berelson, B. (1952). Content analysis in communication research.

Berinsky, A. J. (1999). The two faces of public opinion. *American Journal of Political Science*, 1209-1230.

Black, J. H., Niemi, R. G., & Powell, G. B. (1987). Age, resistance, and political learning in a new environment: The case of Canadian immigrants. *Comparative Politics*, 73-84.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. *Text mining: classification, clustering, and applications*, *10*, 71.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

Bligh, M. C., Kohles, J. C., & Meindl, J. R. (2004). Charting the language of leadership: a methodological investigation of President Bush and the crisis of 9/11. *Journal of Applied Psychology*, *89*(3), 562.

Block, J., & Block, J. H. (2006). Nursery school personality and political orientation two decades later. *Journal of Research in Personality*, *40*(5), 734-749.

Boicu, R. (2011). *Discursive norms in blogging* (No. 41136). University Library of Munich, Germany.

Bollier, D., & Firestone, C. M. (2010). *The promise and peril of big data* (p. 56). Washington, DC, USA: Aspen Institute, Communications and Society Program.

Boyd, D. (2006). A blogger\'s blog: Exploring the definition of a medium. *Reconstruction*, *6*(4).

Branigan, H. P., Pickering, M. J., Pearson, J., & McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, *42*(9), 2355-2368.

Brooks, C., & Manza, J. (1997). The social and ideological bases of middle-class political realignment in the United States, 1972 to 1992. *American Sociological Review*, 191-208.

Carney, D. R., Jost, J. T., Gosling, S. D., & Potter, J. (2008). The secret lives of liberals and conservatives: Personality profiles, interaction styles, and the things they leave behind. *Political Psychology*, *29*(6), 807-840.

Carr, D. (2008) The Media Equation - How Obama tapped into Social Media's Power. *The New York Times.* November 9, 2008: New York, NY.

Castells, M. (1996). The information age: Economy, society, and culture. Volume I: The rise of the network society.

Castells, M. (2011). *The rise of the network society: The information age: Economy, society, and culture* (Vol. 1). John Wiley & Sons.

Chen, J., Hsieh, G., Mahmud, J. U., & Nichols, J. (2014). Understanding individuals' personal values from social media word use. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 405-414). ACM.

Chong, D., & Druckman, J. N. (2007). Framing theory. *Annu. Rev. Polit. Sci.*, *10*, 103-126.

Chou, W. Y. S., Hunt, Y. M., Beckjord, E. B., Moser, R. P., & Hesse, B. W. (2009). Social media use in the United States: implications for health communication. *Journal of medical Internet research*, *11*(4).

Chung, C. J., & Park, H. W. (2010). Textual analysis of a political message: The inaugural addresses of two Korean presidents. *Social science information*, *49*(2), 215-239.

Chung, C. K., & Pennebaker, J. W. (2012). Linguistic Inquiry and Word Count (LIWC): pronounced "Luke",… and other useful facts. *Applied natural language processing and content analysis: Identification, investigation, and resolution. Hershey, PA: IGI Global*, 133-145.

Chung, C., & Pennebaker, J. W. (2007). The psychological functions of function words. *Social communication*, 343-359.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science*, *15*(10), 687-693.

Coleman, S., & Wright, S. (2008). Political blogs and representative democracy. *Information Polity*, *13*(1), 1-6.

Collingwood, L., Barreto, M. A., & Donovan, T. (2012). Early Primaries, Viability and Changing Preferences for Presidential Candidates. *Presidential Studies Quarterly*, *42*(2), 231-255.

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011, October). Predicting the political alignment of twitter users. In *Privacy, Security, Risk* *and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (pp. 192-199). IEEE.

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011, July). Political polarization on twitter. In *ICWSM*.

Conover, P. J., & Feldman, S. (1981). The origins and meaning of liberal/conservative self identification. American Journal of Political Science, 25, 617–645.

Cornfield, M., Carson, J., Kalis, A., & Simon, E. (2005). Buzz, blogs, and beyond: The Internet and the national discourse in the fall of 2004.

Cortina-Borja, M., & Chappas, C. (2006). A stylometric analysis of newspapers, periodicals and news scripts. *Journal of Quantitative Linguistics*, *13*(2-3), 285-312.

Dahlgren, P. (2009). *Media and political engagement*. Cambridge: Cambridge University Press.

Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S. (2011, March). Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web* (pp. 745-754). ACM.

Day, M. V., Fiske, S. T., Downing, E. L., & Trail, T. E. (2014). Shifting liberal and conservative attitudes using moral foundations theory. *Personality and Social Psychology Bulletin*, *40*(12), 1559-1573.

Dehghani, M., Gratch, J., Sachdeva, S., & Sagae, K. (2011). Analyzing Conservative and Liberal Blogs Related to the Construction of the Ground Zero Mosque. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1853-1858).

Dennen, V. P., & Pashnyak, T. G. (2008). Finding community in the comments: The role of reader and blogger responses in a weblog community of practice. *International Journal of Web Based Communities*, *4*(3), 272-283.

Dovidio, J. F., Brown, C. E., Heltman, K., Ellyson, S. L., & Keating, C. F. (1988). Power displays between women and men in discussions of gender-linked tasks: A multichannel study. *Journal of personality and Social Psychology*, *55*(4), 580.

Dreifus, C. (1986). Gore Vidal: An Interview. *The Progressive*. Vol. 50, No. 9. September 1986.

Drezner, D., & Farrell, H. (2004). The power and politics of blogs. American Political Science Association.

Du, H. S., & Wagner, C. (2006). Weblog success: Exploring the role of technology. *International Journal of Human-Computer Studies*, *64*(9), 789-798.

Duggan, M., & Smith, A. (2013). Social media update 2013. *Pew Internet and American Life Project*.

Dunleavy, P. (1992). Democracy, bureaucracy and public choice. *Public administration*, *95*.

Dunleavy, P. (2014). *Democracy, bureaucracy and public choice: economic approaches in political science*. Routledge.

Dychtwald, K. (1999). *Age power: How the 21st century will be ruled by the new old*. New York: Jeremy P. Tarcher/Putnam.

Eckert, P., & McConnell-Ginet, S. (2003). *Language and gender*. Cambridge University Press.

Entman, R. M. (2010). Media framing biases and political power: Explaining slant in news of Campaign 2008. *Journalism*, *11*(4), 389-408.

Erikson R. S., Tedin K. L. (2003). *American Public Opinion*. New York: Longman. 6th ed.

Eveland, W. P., & Dylko, I. (2007). Reading political blogs during the 2004 election campaign: Correlates and political consequences. *Blogging, citizenship, and the future of media*, 105-126.

Fine, T. S. (1992). The impact of issue framing on public opinion: Toward affirmative action programs. *The Social Science Journal*, *29*(3), 323-334.

Fischer, U., McDonnell, L., & Orasanu, J. (2007). Linguistic correlates of team performance: Toward a tool for monitoring team functioning during space missions. *Aviation, space, and environmental medicine*, *78*(Supplement 1), B86-B95.

Flesch, R. F. (1951). *How to test readability*. New York: Harper and Brothers.

Forest, B. (2005). The changing demographic, legal, and technological contexts of political representation. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15331-15336.

Freeman, R., & McElhinny, B. (1996). Language and gender. *Sociolinguistics and language teaching*, 218-280.

Furnham, A. (1985). Why do people save? Attitudes to, and habits of saving money in Britain. *Journal of Applied Social Psychology*, *15*(5), 354-373.

Furnham, A., & Argyle, M. (1998). *The psychology of money*. Psychology Press.

Galtung, J., & Ruge, M. H. (1965). The structure of foreign news the presentation of the

Congo, Cuba and Cyprus Crises in four Norwegian newspapers. *Journal of peace research*, *2*(1), 64-90.

Garton, L., Haythornthwaite, C., & Wellman, B. (1997). Studying online social networks. *Journal of Computer-Mediated Communication*, *3*(1), 0-0.

Gayo-Avello, D. (2012). No, you cannot predict elections with twitter. *Internet Computing, IEEE*, *16*(6), 91-94.

Gelman, A. (2009). *Red state, blue state, rich state, poor state: why Americans vote the way they do*. Princeton University Press.

Gerring, J. (1997). Ideology: A definitional analysis. *Political Research Quarterly*, 957-994.

Gilbert, E., & Karahalios, K. (2010, May). Widespread Worry and the Stock Market. In *ICWSM* (pp. 59-65).

Giles, H., & Powesland, P. F. (1975). *Speech style and social evaluation*. Academic Press.

Giles, H., Coupland, N., & COUPLAND, I. (1991). 1. Accommodation theory: Communication, context, and. *Contexts of accommodation: Developments in applied sociolinguistics*, *1*.

Gill, K. E. (2004, May). How can we measure the influence of the blogosphere. In *Proceedings of the WWW'04: workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*.

Gilligan, C. (1977). In a different voice: Women's conceptions of self and of morality. *Harvard educational review*, *47*(4), 481-517.

Gilligan, C., & Attanucci, J. (1988). Two moral orientations: Gender differences and similarities. *Merrill-Palmer Quarterly (1982-)*, 223-237.

Gladwell, M. (2010). Small change. *The New Yorker*, *4*(2010), 42-49.

Gladwell, M. (2011). From innovation to revolution-do social media made protests possible: An absence of evidence. *Foreign Aff.*, *90*, 153.

Glass, J., Bengtson, V. L., & Dunham, C. C. (1986). Attitude similarity in three-generation families: Socialization, status inheritance, or reciprocal influence?. *American Sociological Review*, 685-698.

Gleser GC, Gottschalk LA, Watkins J. 1959. The relationship of sex and intelligence to choice of words: a normative study of verbal behavior. *J. Clin. Psychol.* 15:183–91

Goffman, E. (1959). The presentation of self in everyday life. *Garden City, NY: Anchor*.

Golan, G. J. (2014). Agenda Setting in a 2.0 World: New Agendas in Communication. *Journal of Broadcasting & Electronic Media*, *58*(3), 476-477.

Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2009). Language style matching as a predictor of social dynamics in small groups. *Communication Research*.

Goode, J., & Robinson, J. D. (2013). Linguistic synchrony in parasocial interaction. *Communication Studies*, *64*(4), 453-466.

Gortner, E. M., & Pennebaker, J. W. (2003). The archival anatomy of a disaster: Media coverage and community-wide health effects of the Texas A&M bonfire tragedy. *Journal of Social and Clinical Psychology*, *22*(5), 580-603.

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*(5), 1029.

Green, D., Palmquist, B., & Schickler, E. (2002). Partisan hearts and minds.

Gumperz, J. J. (1964). Linguistic and Social Interaction in Two Communities1. *American anthropologist*, *66*(6_PART2), 137-153.

Gurzick, D., & Lutters, W. G. (2006, April). From the personal to the profound: understanding the blog life cycle. In *CHI'06 extended abstracts on Human factors in computing systems* (pp. 827-832). ACM.

Hafner, K. (2004) For Some, the Blogging Never Stops. *New York Times,* 27 May 2004.

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*(5827), 998-1002.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20*(1), 98-116.

Haidt, J., Graham, J., & Joseph, C. (2009). Above and below left–right: Ideological narratives and moral foundations. *Psychological Inquiry*, *20*(2-3), 110-119.

Hamlin, A., & Jennings, C. (2011). Expressive political behaviour: Foundations, scope and implications. *British Journal of Political Science*, *41*(03), 645-670.

Hargittai, E., Gallo, J., & Kane, M. (2008). Cross-ideological discussions among conservative and liberal bloggers. *Public Choice*, *134*(1-2), 67-86.

Harp, D., & Tremayne, M. (2006). The gendered blogosphere: Examining inequality using network and feminist theory. *Journalism & Mass Communication Quarterly*, *83*(2), 247-264.

Hayden, S. (2003). Family metaphors and the nation: Promoting a politics of care through the Million Mom March. *Quarterly Journal of Speech*, *89*(3), 196-215.

Hermann, M. G. (1980). Assessing the personalities of Soviet Politburo members. *Personality and Social Psychology Bulletin*, *6*(3), 332-352.

Herring, S. C., Scheidt, L. A., Bonus, S., & Wright, E. (2004, January). Bridging the gap: A genre analysis of weblogs. In *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on* (pp. 11-pp). IEEE.

Herring, S. C., Scheidt, L. A., Wright, E., & Bonus, S. (2005). Weblogs as a bridging genre. *Information Technology & People*, *18*(2), 142-171.

Hirsh, J. B., DeYoung, C. G., Xu, X., & Peterson, J. B. (2010). Compassionate liberals and polite conservatives: Associations of agreeableness with political ideology and moral values. *Personality and Social Psychology Bulletin*, *36*(5), 655-664.

Holmes, D. I. (1992). A stylometric analysis of Mormon scripture and related texts. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 91-120.

Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and linguistic computing*, *13*(3), 111-117.

Holmes, J., & Meyerhoff, M. (Eds.). (2008). *The handbook of language and gender* (Vol. 25). John Wiley & Sons.

Holsti, O. R. (1968). Content Analysis. In *The Handbook of Social Psychology, Vol. 2*, eds. G. Lindzey and E. Aronson, Reading, MA: Addison-Wesley.

Horowitz, M., Wilner, N., & Alvarez, W. (1979). Impact of Event Scale: a measure of subjective stress. *Psychosomatic medicine*, *41*(3), 209-218.

Huffaker, D. A., & Calvert, S. L. (2005). Gender, identity, and language use in teenage blogs. *Journal of Computer-Mediated Communication*, *10*(2), 00-00.

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*(3), 549.

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.

Iyengar, S. (1990). Framing responsibility for political issues: The case of poverty. *Political behavior*, *12*(1), 19-40.

Iyengar, S. (2005, November). Speaking of values: The framing of American politics. In *The Forum* (Vol. 3, No. 3).

Iyengar, S., & Simon, A. (1993). News coverage of the gulf crisis and public opinion a study of agenda-setting, priming, and framing. *Communication research*, *20*(3), 365-383.

Jeffries, J. W. (1990). The" New" New Deal: FDR and American Liberalism, 1937-1945. *Political Science Quarterly*, 397-418.

Jelen, T. G., Thomas, S., & Wilcox, C. (1994). The gender gap in comparative perspective. *European Journal of Political Research*, *25*(2), 171-186.

Johnson, T. J., Kaye, B. K., Bichard, S. L., & Wong, W. J. (2007). Every Blog Has Its Day: Politically‐interested Internet Users' Perceptions of Blog Credibility. *Journal of Computer‐Mediated Communication*, *13*(1), 100-122.

Jost, J. T., Federico, C. M., & Napier, J. L. (2009). Political ideology: Its structure, functions, and elective affinities. *Annual review of psychology*, *60*, 307-337.

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological bulletin*, *129*(3), 339.

Jost, J. T., Nosek, B. A., & Gosling, S. D. (2008). Ideology: Its resurgence in social, personality, and political psychology. *Perspectives on Psychological Science*, *3*(2), 126-136.

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*, 0261927X13502654.

Kaefer, F., Heilman, C. M., & Ramenofsky, S. D. (2005). A neural network application to consumer classification to improve the timing of direct marketing activities. *Computers & Operations Research*, *32*(10), 2595-2615.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, *53*(1), 59-68.

Kassarjian, H. H. (1977). Content analysis in consumer research. *Journal of consumer research*, 8-18.

Kennedy, E. (1979). " Ideology" from Destutt De Tracy to Marx. *Journal of the History of Ideas*, 353-368.

Kerlinger, F. H. (1964). *Foundations of Behavioral Research: Educational and Psychological Inquiry.* New York: Holt, Rinehart & Winston.

Khan, R., Misra, K., & Singh, V. (2013). Ideology and brand consumption. *Psychological science*, 0956797612457379.

Kinder D. R., (1998). "Opinion and Action in the Realm of Politics." in Daniel Gilbert, Susan Fiske, and Gardner Lindsey (eds.) *Handbook of Social Psychology*, fourth edition. Boston: McGraw Hill

Knight, K. (2006). Transformations of the Concept of Ideology in the Twentieth Century. *American Political Science Review*, *100*(4), 619.

Knutson, K. M., Wood, J. N., Spampinato, M. V., & Grafman, J. (2006). Politics on the brain: An fMRI investigation. *Social Neuroscience,*, *1*(1), 25-40.

Kohlberg, L. (1981). The philosophy of moral development: Moral stages and the idea of justice.

Kohlberg, L. (1984). The psychology of moral development: Essays on moral development. San Francisco: Harper and Row.

Koop, R., & Jansen, H. J. (2009). Political blogs and blogrolls in Canada: forums for democratic deliberation?. *Social Science Computer Review*.

Koppel, M., & Schler, J. (2003, August). Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis* (Vol. 69, pp. 72-80).

Koppel, M., Akiva, N., Alshech, E., & Bar, K. (2009). Automatically classifying documents by ideological and organizational affiliation.

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, *17*(4), 401-412.

Labov, W. (1964). *The social stratification of English in New York City* (Doctoral dissertation, Columbia university.).

Labov, W. (1972). *Sociolinguistic patterns* (No. 4). University of Pennsylvania Press.

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language variation and change*, *2*(02), 205-254.

Lakoff, G. (2008). *The political mind: A cognitive scientist's guide to your brain and its politics*. Penguin.

Lakoff, G. (2010). *Moral politics: How liberals and conservatives think*. University of Chicago Press.

Lakoff, R. T. (1972). *Language and woman's place*. University of Michigan, Center for Advance Study in the Behavioral Sciences.

Larsson, A. O., & Moe, H. (2012). Studying political microblogging: Twitter users in the 2010 Swedish election campaign. *New Media & Society*, *14*(5), 729-747.

Lawrence, E., Sides, J., & Farrell, H. (2010). Self-segregation or deliberation? Blog readership, participation, and polarization in American politics. *Perspectives on Politics*, *8*(01), 141-157.

Levenson, H., & Miller, J. (1976). Multidimensional locus of control in sociopolitical activists of conservative and liberal ideologies. *Journal of personality and social psychology*, *33*(2), 199.

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational behavior and human decision processes*, *76*(2), 149-188.

Livne, A., Simmons, M. P., Adar, E., & Adamic, L. A. (2011). The Party Is Over Here: Structure and Content in the 2010 Election. *ICWSM*, *11*, 17-21.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., & Pearce, I. (2011). The Arab Spring| the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International journal of communication*, *5*, 31.

Lublin, D. (1999). *The paradox of representation: Racial gerrymandering and minority interests in Congress*. Princeton University Press.

Lucas, C., Nielsen, R., Roberts, M., Stewart, B., Storer, A., & Tingley, D. (2013). *Computer assisted text analysis for comparative politics*. Working paper.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., ... & McKinsey Global Institute. (2011). Big data: The next frontier for innovation, competition, and productivity.

Mark, G., Bagdouri, M., Palen, L., Martin, J., Al-Ani, B., & Anderson, K. (2012, February). Blogs as a collective war diary. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 37-46). ACM.

Marres, N., & Weltevrede, E. (2013). Scraping the Social? Issues in live social research. *Journal of Cultural Economy*, *6*(3), 313-335.

Martin, J. R. (2004). Mourning: how we get aligned. *Discourse & Society*, *15*(2-3), 321-344.

Matthews, R. A., & Merriam, T. V. (1993). Neural computation in stylometry I: An application to the works of Shakespeare and Fletcher. *Literary and Linguistic Computing*, *8*(4), 203-209.

McAdams, D. P., Albaugh, M., Farber, E., Daniels, J., Logan, R. L., & Olson, B. (2008). Family metaphors and moral intuitions: how conservatives and liberals narrate their lives. *Journal of personality and social psychology*, *95*(4), 978.

McCarty, N., Poole, K. T., & Rosenthal, H. (2006). *Polarized America: The dance of ideology and unequal riches* (Vol. 5). mit Press.

McCarty, N., Poole, K. T., & Rosenthal, H. (2006). *Polarized America: The dance of ideology and unequal riches* (Vol. 5). mit Press.

McClosky, H. (1964). Consensus and ideology in American politics. *American Political Science Review*, *58*(02), 361-382.

McKenna, L., & Pole, A. (2008). What do bloggers do: an average day on an average political blog. *Public Choice*, *134*(1-2), 97-108.

Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: a psychometric analysis of students' daily social environments and natural conversations. *Journal of personality and social psychology*, *84*(4), 857.

Meraz, S. (2009). The Many Faced 'You'of Social Media. *Journalism and Citizenship: New Agendas in Communication*, 123-48.

Meraz, S. (2011). Using time series analysis to measure intermedia agenda-setting influence in traditional media and political blog networks. *Journalism & Mass Communication Quarterly*, *88*(1), 176-194.

Merriam, T. V., & Matthews, R. A. (1994). Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe. *Literary and Linguistic Computing*, *9*(1), 1-6.

Miller, D. T. (1999). The norm of self-interest. *American Psychologist, 54,* 1053–1060.

Mishne, G., & Glance, N. S. (2006, March). Predicting Movie Sales from Blogger Sentiment. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (pp. 155-158).

Mitchell, A., Gottfried, J., Kiley, J., Matsa, K. (2014). Political Polarization and Media Habits. Pew Research Center. Retrieved 11/5/2014 from http://www.journalism.org/2014/10/21/political-polarization-media-habits/

Mitra, R. (2010). Resisting the spectacle of pride: queer Indian bloggers as interpretive communities. *Journal of Broadcasting & Electronic Media*, *54*(1), 163-178.

Moran, M., Seaman, J., & Tinti-Kane, H. (2011). Teaching, Learning, and Sharing: How Today's Higher Education Faculty Use Social Media. *Babson Survey Research Group*.

Morin, R. (2012) Rising Share of Americans See Conflict Between Rich and Poor. Pew Research Center. Accessed November 26, 2014: http://www.pewsocialtrends.org/2012/01/11/rising-share-of-americans-see-conflict-between-rich-and-poor/

Motyl, M. (2012). Party Evolutions in Moral Intuitions: A Text-Analysis of US Political Party Platforms from 1856-2008. *Available at SSRN 2158893*.

Mullen, T., & Malouf, R. (2006). A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (pp. 159-162).

Muller, M., Ehrlich, K., Matthews, T., Perer, A., Ronen, I., & Guy, I. (2012, May). Diversity among enterprise online communities: collaborating, teaming, and innovating through social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2815-2824). ACM.

Nardi, B. A., Schiano, D. J., & Gumbrecht, M. (2004, November). Blogging as social activity, or, would you let 900 million people read your diary?. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work* (pp. 222-231). ACM.

Nardi, B. A., Schiano, D. J., Gumbrecht, M., & Swartz, L. (2004). Why we blog. *Communications of the ACM*, *47*(12), 41-46.

Nelson, T. E., Oxley, Z. M., & Clawson, R. A. (1997). Toward a psychology of framing effects. *Political behavior*, *19*(3), 221-246.

Newman, D. R., Webb, B., & Cochrane, C. (1995). A content analysis method to measure critical thinking in face-to-face and computer supported group learning. *Interpersonal Computing and Technology*, *3*(2), 56-77.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577-8582.

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*(3), 211-236.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, *29*(5), 665-675.

Nguyen, D., & Rosé, C. P. (2011, June). Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Languages in Social Media* (pp. 76-85). Association for Computational Linguistics.

Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, *21*(4), 337-360.

Nielsen. (2011). State of the media: the social media report. *NM Incite, A Neilsen/A Mckinsey Company*.

Norton, M. I., & Ariely, D. (2011). Building a better America—One wealth quintile at a time. *Perspectives on Psychological Science*, *6*(1), 9-12.

Nunberg, G. (2007). *Talking right: How conservatives turned liberalism into a tax-raising, latte-drinking, sushi-eating, volvo-driving… freak show*. PublicAffairs.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, *11*, 122-129.

Olson, J., Ouyang, Y., Poe, J., Trantham, A., & Waterman, R. W. (2012). The teleprompter presidency: Comparing Obama's campaign and governing rhetoric. *Social Science Quarterly*, *93*(5), 1402-1423.

Page, B. I., & Jones, C. C. (1979). Reciprocal effects of policy preferences, party loyalties and the vote. *American Political Science Review*, *73*(04), 1071-1089.

Page, B. I., Bartels, L. M., & Seawright, J. (2013). Democracy and the policy preferences of wealthy Americans. *Perspectives on Politics*, *11*(01), 51-73.

Paisley, W. J. (1969). Studying style as deviation from encoding norms. *The analysis of communication contents: Developments in scientific theories and computer techniques*, 4458.

Parsons T. (1951). *The Social System*. New York: Free Press

Pennebaker, J. W. (1993). Putting stress into words: Health, linguistic, and therapeutic implications. *Behaviour research and therapy*, *31*(6), 539-548.

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

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, *77*(6), 1296.

Pennebaker, J. W., & Lay, T. C. (2002). Language use and personality during crises: Analyses of Mayor Rudolph Giuliani's press conferences. *Journal of Research in Personality*, *36*(3), 271-282.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*, 2001.

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*(1), 172.

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*(1), 547-577.

Perlmutter, D. (2008) *Blogwars*. Oxford University Press.

Phillips, J. R. (1973). Syntax and vocabulary of mothers' speech to young children: Age and sex comparisons. *Child development*, 182-185.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, *27*(02), 169-190.

Radday, Y. and Wickman, D. (1975). The Unity of Zechariah in the light of statistical linguistics, Zeit. für Alttestamentliche Wissenschaft, 87, 30–55.

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: moral motives for unity, hierarchy, equality, and proportionality. *Psychological review*, *118*(1), 57.

Rall, T. (2005) WHO WATCHES THE WATCHDOGS? Bloggers and the New McCarthyism. March 2, 2005. Santa Fe Reporter.

Rhodebeck, L. A. (1993). The politics of greed? Political preferences among the elderly. *The Journal of Politics*, *55*(02), 342-364.

Rindermann, H., Flores-Mendoza, C., & Woodley, M. A. (2012). Political orientations, intelligence and education. *Intelligence*, *40*(2), 217-225.

Rogan, R. G. (2011). Linguistic style matching in crisis negotiations: A Comparative analysis of suicidal and surrender outcomes. *Journal of police crisis negotiations*, *11*(1), 20-39.

Rosenstiel, T., & Mitchell, A. (2012). How the presidential candidates use the web and social media. Pew Research Center's Project for Excellence in Journalism. Retrieved July 15, 2014, from http://www.journalism.org/2012/08/15/how-presidential-candidates-use-web-and-social-media/

Rude, S., Gortner, E. M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, *18*(8), 1121-1133.

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

Sang, E. T. K., & Bos, J. (2012, April). Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media* (pp.

53-60). Association for Computational Linguistics.

Sarno, D. (2008) Obama, the Social Media President. *The Los Angeles Times*. November 18, 2008: Los Angeles, CA.

Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology*, *41*(5), 885-899.

Schneider, G. L. (2009). *The conservative century: from reaction to revolution*. Rowman & Littlefield.

Schnurr, P. P., Rosenberg, S. D., Oxman, T. E., & Tucker, G. J. (1986). A methodological note on content analysis: Estimates of reliability. *Journal of personality assessment*, *50*(4), 601-609.

Schultz, C. (1996). Polarization and inefficient policies. *The Review of Economic Studies*, *63*(2), 331-344.'

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, *8*(9), e73791.

Scott, E. (2004). 'Big Media'Meets the 'Bloggers': Coverage of Trent Lott's Remarks at Strom Thurmond's Birthday Party. *Kennedy School of Government Case Program. Case*, *1731*.

Sears, D. O., & Funk, C. L. (1999). Evidence of the long-term persistence of adults' political predispositions. *The Journal of Politics*, *61*(01), 1-28.

Shachtman, N. (2002). Blogs make the headlines. *Wired News*, December 23, 2002.

Shapiro, I. (2012). *The moral foundations of politics*. Yale University Press.

Shapiro, R. Y., & Mahajan, H. (1986). Gender Differences in Policy Preferences: A Summary of Trends from the 1960s to the 1980s. *Public Opinion Quarterly*, *50*(1), 42-61.

Shaw, A., & Benkler, Y. (2012). A tale of two blogospheres discursive practices on the left and right. *American Behavioral Scientist*, *56*(4), 459-487.

Shaw, D. R. (1999). A study of presidential campaign event effects from 1952 to 1992. *The Journal of Politics*, *61*(02), 387-422.

Shirky, C. (2003). Power laws, weblogs, and inequality. *Clay Shirky's writings about the Internet*, *8*.

Shirky, C. (2011). The political power of social media: Technology, the public sphere, and political change. *Foreign affairs*, 28-41.

Simon, H. A. (1995). Rationality in political behavior. *Political Psychology* 16: 45-61.

Slatcher, R. B., Chung, C. K., Pennebaker, J. W., & Stone, L. D. (2007). Winning words: Individual differences in linguistic style among US presidential and vice presidential candidates. *Journal of Research in Personality*, *41*(1), 63-75.

Spiegelman, M., Terwilliger, C., & Fearing, F. (1952). The content of comic strips: A study of a mass medium of communication. *The Journal of Social Psychology*, *35*(1), 37-57.

Stieglitz, S., & Dang-Xuan, L. (2012). Political communication and influence through microblogging--An empirical analysis of sentiment in Twitter messages and retweet behavior. In *System Science (HICSS), 2012 45th Hawaii International Conference on* (pp. 3500-3509). IEEE.

Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, *63*(4), 517-522.

Stone, G. C., & McCombs, M. E. (1981). Tracing the time lag in agenda-setting. *Journalism Quarterly*, *58*(1), 51-55.

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The General Inquirer: A Computer Approach to Content Analysis.

Suchman, L. (1995). Making work visible. *Communications of the ACM*, *38*(9), 56-ff.

Sunstein, C. R. (2003). *Republic.com*. Princeton, NJ: Princeton University Press.

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*(1), 24-54.

Tausczik, Y. R., & Pennebaker, J. W. (2013, April). Improving teamwork using real-time language feedback. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 459-468). ACM.

Taylor, P. J., & Thomas, S. (2008). Linguistic style matching and negotiation outcome. *Negotiation and Conflict Management Research*, *1*(3), 263-281.

Technorati. List of Top Political Blogs. http://www.technorati.com/political/. Accessed 18 September 2013.

Thompson, J. B. (2013). *Ideology and modern culture: Critical social theory in the era of mass communication*. John Wiley & Sons.

Thomson, R., Murachver, T., & Green, J. (2001). Where is the gender in gendered language?. *Psychological Science*, *12*(2), 171-175.

Tremayne, M. (Ed.). (2012). *Blogging, citizenship, and the future of media*. Routledge.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM*, *10*, 178-185.

Vaisey, S., & Miles, A. (2014). Tools from moral psychology for measuring personal moral culture. *Theory and Society*, *43*(3-4), 311-332.

Vambheim, S. M., Wangberg, S. C., Johnsen, J. A. K., & Wynn, R. (2012). Language use in an internet support group for smoking cessation: development of sense of community. *Informatics for Health and Social Care*, *38*(1), 67-78.

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010, April). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1079-1088). ACM.

Wagner, M., & Kritzinger, S. (2012). Ideological dimensions and vote choice: Age group differences in Austria. *Electoral Studies*, *31*(2), 285-296.

Walker, D. M. (2006). Blog commenting: A new political information space. *Proceedings of the American Society for Information Science and Technology*, *43*(1), 1-10.

Walker, L. J. (2006). Gender and morality. *Handbook of moral development*, 93-115.

Wallsten, K. (2007). Agenda setting and the blogosphere: An analysis of the relationship between mainstream media and political blogs. *Review of Policy Research*, *24*(6), 567-587.

Wang, H. C., & Fussell, S. (2010, February). Groups in groups: conversational similarity in online multicultural multiparty brainstorming. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 351-360). ACM.

Warren, T., & Rayner, K. (2004). Top-down influences in the interactive alignment model: The power of the situation model. *Behavioral and Brain Sciences*, *27*(02), 211-211.

Weintraub, W. (1986). Personality profiles of American presidents as revealed in their public statements: The presidential news conferences of Jimmy Carter and Ronald Reagan. *Political Psychology*, 285-295.

West, C., & Zimmerman, D. H. (1987). Doing gender. *Gender & society*, *1*(2), 125-151.

Whissell, C. (1996). Traditional and emotional stylometric analysis of the songs of Beatles Paul McCartney and John Lennon. *Computers and the Humanities*, *30*(3), 257-265.

Whissell, C., & Sigelman, L. (2001). The times and the man as predictors of emotion and style in the inaugural addresses of US presidents. *Computers and the Humanities*, *35*(3), 255-272.

Williams, C., & Gulati, G. (2008). What is a social network worth? Facebook and vote share in the 2008 presidential primaries. American Political Science Association.

Wojcik, S., Hovasapian, A., Graham, J., Motyl, M., Ditto, P. Conservatives report, but liberals display, greater happiness. *Science* 13 March 2015: Vol. 347 no. 6227 pp. 1243-1246

Woodly, D. (2008). New competencies in democratic communication? Blogs, agenda setting and political participation. *Public Choice*, *134*(1-2), 109-123.

Xenos, M. (2008). New mediated deliberation: Blog and press coverage of the Alito nomination. *Journal of Computer-Mediated Communication*, *13*(2), 485-503.

Yano, T., Resnik, P., & Smith, N. A. (2010). Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 152-158). Association for Computational Linguistics.

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*(3), 363-373.

Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on* (pp. 427-434). IEEE.

Zamboni, G., Gozzi, M., Krueger, F., Duhamel, J. R., Sirigu, A., & Grafman, J. (2009). Individualism, conservatism, and radicalism as criteria for processing political beliefs: a parametric fMRI study. *Social Neuroscience*, *4*(5), 367-383.

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through twitter "I hope it is not as bad as I fear". *Procedia-Social and Behavioral Sciences*, *26*, 55-62.

Zhao, D., & Rosson, M. B. (2009). How and why people Twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work* (pp. 243-252). ACM.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological bulletin*, *123*(2), 162.