Automated Extraction of Mental Models from Obama's and McCain's Campaign Speeches: A Natural Language Processing and Statistical Approach

Peter Muhlberger, Ph.D.

Director, Center for Communication Research
College of Mass Communications
Texas Tech University
Lubbock, TX 79409
Phone: 806-742-6500
Fax: 806-742-1085
E-mail: pmuhl1848@gmail.com

Weiwu Zhang, Ph.D.

Department of Public Relations
College of Mass Communications
Texas Tech University
Lubbock, TX 79409
Phone: 806-742-6500
Fax: 806-742-1085
E-mail: weiwu.zhang@ttu.edu

Automated Extraction of Mental Models from Obama's and McCain's Campaign Speeches: A Natural Language Processing and Statistical Approach

Abstract

This paper applies natural language processing (NLP) and statistical analysis to speeches of John McCain and Barack Obama in a part of the 2008 presidential race in an effort to extract the candidates' ideological mental models. A focus of the paper is to test moderately simple methods that help identify, from the over 6000 unique words utilized by the candidates, a statistically manageable subset that clarifies conceptual differences. NLP was applied to identify the parts of speech for each word in the candidates' speeches. This information helps disambiguate word meaning and narrow the choice of words. Statistical bootstrapping and a simple Bayesian technique provided two key indicators to further narrow the selection of words to those that involved statistically significant differences between the candidates and that also were likely to play a substantial role in the candidates' conceptual systems. Findings reveal highly significant differences between the candidates in their choice of words. Identified terms and their use in context suggest that McCain embraces a touch of free market ideology, but, far more prominently, a conservative form of the political philosophy of republicanism—one that values the state and political institutions above the individual. McCain's language differentially stresses lofty goals, abstract policies, institutional authorities, and national unity. He presents himself more formally and does not appear to offer explanations in his speeches at the same rate as Obama. Obama, in contrast, presents himself as a grounded populist focused on everyday

concerns and people. His language differentially stresses everyday issues, concrete policies, and ordinary people as well as far more freely examining social divisions. He presents himself more informally and offers appreciably more explanations. Contrary to some pundit wisdom, Obama's language, at least, appears to be populist rather than elitist. Multivariate statistical results indicate that the terms identified by the techniques deployed are moderately powerful predictors of which candidate is speaking, even on separate data than that used to construct the model. This suggests the analysis may have captured an important part of what makes these candidates different.

Introduction

Digitalization makes huge volume of communication content readily available to social scientists. Increasingly as database creation has become easier and software development continues, content analysts use computers to identify content, access content, and code content. In addition to its flexibility and potential to complete certain tasks with greater speed, computer content analysis is particularly helpful for some research projects (Holsti, 1969): (1) when the unit of analysis is the symbol or word, and analysis concerns number of times a symbol or word is used; (2) when the analysis is extremely complex, for instance, using a large number of categories with a large number of recording units, such as when inference is to be based on the co-occurrence of two or more terms in the same sentence; (3) when the analysis involves examining the data in multiple ways. Computers allow complicated manipulations to make nuanced understanding of text data possible; and (4) when the data come from documents that are crucial to a variety of disciplines, researchers, and lines of inquiry, and might be used in multiple studies.

While progress in data accessibility provides a lot of opportunities for social scientists to do computer content analysis, it also presents many challenges. One central challenge of working with massive data is that it must be culled, organized, classified, and summarized so that the researcher can use it for analysis. Another challenge for text annotation is the extent to which humans are involved in the annotation process. Humans do all the annotation in manual methods; while unsupervised (computer-based) learning algorithms detect patterns in text and require no/little human involvement. Somewhere in between are supervised learning algorithms that use small amount of manually annotated training data to validate large amount of annotations by the

automated systems (Cardi and Wilkerson, 2008).

Each approach will use different techniques to handle different tasks with its pros and cons. Unsupervised learning algorithms can be implemented with greater flexibility and speed but may not be reliable or valid; pure manual method is generally slow to apply and time-consuming but remains the most reliable and valid option. Supervised learning algorithms seek to strike a balance between the speed of unsupervised method and reliability and validity benefits of the manual method (e.g., Russell & Norvig, 2002).

This paper presents a preliminary exploration of the feasibility and value of extracting ideological mental models from textual data, in this case the speeches of the presidential candidates Barack Obama and John McCain. It presents an approach involving natural language processing and statistical analysis that makes an advance toward extracting such models. We define a "mental model" as the key concepts and the conditional relationships between them that a person brings to bear in a given context, such as the context of a candidate in a presidential campaign.

Scholars of public opinion and media may be interested in the automated extraction of mental models for several reasons. The media, including new media such as the internet, offer researchers enormous volumes of textual data. Content analysis of such data using human coding is extremely time consuming, may pick up on patterns that are not robust across larger bodies of text, and are subject to the vicissitudes of subjective interpretation. Automated analysis of meaning is imperfect as well, failing to fully capture the complex and subtle meanings of language. Nevertheless, such analysis may yield insights that are not feasible with content analysis, both by detecting statistical patterns that would be undetectable to human coders and establishing with statistical certainty that certain word use and connections between

words used explain differences, such as between liberal and conservative speech. Also, automated analysis can highlight differences that can then be subject to human content coding, greatly reducing the effort required to extract meaning from text. The statistical approach for creating models of word use developed in this paper can be trained on one set of data, say Obama's and McCain's speeches, and then applied elsewhere, such as on the news media, to see how well the mental models in the speeches are reflected in the media. The prevalence of these models in the media might then be used to explain candidate popularity. These tools, then, can be used to analyze relationships between the meanings embedded in different information sources as well as political outcomes.

The first step toward developing methods of extracting mental models from text must address the profusion of linguistic possibilities. The mental models described here might be analyzed using Bayesian networks, which clarify the effect of the presence of words in the network on the probability of the occurrence of any given word. The computations necessary for such an analysis rises super-exponentially with an increase in the number of terms. This makes impossible analyzing ordinary speech without some process by which to select a subset of words. People in ordinary speech deploy thousands of words and most of these words occur very rarely, even without taking into consideration that the same word can have multiple meanings. Thus, this paper will introduce our efforts to extract mental models by focusing on natural language processing and statistical methods of winnowing the field of words to a manageable number. Also, our results here are preliminary. Nevertheless, they are quite encouraging. They show that the deployed methods suggest insights about the 2008 campaigns and the apparent thought patterns of the two major party presidential candidates. Specifically, the models indicate that John McCain embraces a conservative republican political philosophy,

while Obama stresses equality and concrete rather than lofty concerns. Also, a model developed with training data significantly and appreciably predicts who the speaker is in randomly selected non-training data, indicating that the model does capture real differences between speakers.

Literature Review

"Where you stand depends on where you sit." Ideology as a system of beliefs guides individuals' opinions on given issues. An ideology expresses a view of which issue positions go together, the "knowledge of what-goes-with-what" (Poole, 2003, p.3). Because ideology constrains each person's views on issues, such influences will be identifiably different for liberals and conservatives (Yu, Kaufmann, and Diermeier, 2008).

A number of studies have tackled extracting political figures' ideology from their partisan speeches using a variety of methods. Coffey (2005) used computer content analysis program TEXTPACK to directly measure governors' ideology by analyzing their State of the State addresses in 2000 and 2001. TEXTPACK program, originally designed for the analysis of open-ended survey responses, has been broadened to include most of content analysis. It produces word frequencies, alphabetical lists, key word in context and key word out of context searches, word comparisons between two texts, and automatic coding based on user-created dictionaries (Neuendorf, 2002). The ideological content of these speeches also distinguishes governors based on their party affiliation.

However, this traditional computer-coded content analysis merely reproduces the hand-coding of texts using algorithms to match texts to coding dictionaries. With proper dictionaries matching word or phrases to specific ideology, traditional computer-coding of texts can generate more valid estimate of ideology than hand-coded content analyses of the same texts (e.g., Laver

& Garry, 2000). But this approach still needs heavy human involvement in the generation of coding dictionaries that are sensitive to a particular context. Because such dictionary generation is time-consuming and costly, the temptation is go for large-generic dictionaries, which can be insensitive to context with much bias from human coders (e.g., Laver, Benoit, and Garry, 2003). Laver et al (2003) presented a new way of extracting policy positions from political texts that treats texts not as discourses to be understood but as data containing information about the positions of the texts' authors. In this way, analysts can estimate policy positions in any language. This in turn allows analysts to make informed judgment of the extent to which differences between two estimated policy positions have substantive difference or merely as products of measurement error.

Recent research characterized and compared the information-processing styles of political leaders from annotating political speeches. Grounded in theory of treating what leaders say as indicative of how they think, Dyson (2008) used a text analysis protocol of high-vs. low-complexity words to apply to the universe of British Prime Ministers' responses to foreign policy questions in the House of Commons from 1945-2008. The results showed that the techniques can discriminate reliably leaders with differing levels of cognitive complexity.

Yu, Kaufmann, and Diermeier (2008) used the standard process for training and testing text categorization systems (supervised learning algorithms, in general) to make binary document-level decisions to classify U.S. Congressional speeches based on the political party affiliation of the speaker (i.e., Democrat or Republican). Specifically, they tested the person-dependency and time-dependency of ideology classifiers trained on a variety of Congressional speech subsets and found that the ideology classifiers trained on the 2005 House speeches can be generalized to the Senate speeches of the same year, but not vice versa, in keeping with the

expectation that the House is more partisan than the Senate.

Method

Data

For purposes of this analysis, we utilized the entire text of 71 public speeches by John McCain and Barack Obama from January 8, 2008 to July 10, 2008 that were available on the candidates' websites.

Procedures

The methods in this paper are oriented to extracting meaning from text by identifying significant differences in how people use words and, ultimately, concepts. Of course, many of the words people use are merely linguistic scaffolding and therefore do not correspond to concepts. Also, the same word can be used with different meaning. The word "account," for example, can signify a bank account, an account of an event, whether a person can account for an action, or a "Millennium Account." Notice that these uses of the word correspond to different parts of speech, respectively: noun, noun with a different meaning, verb, and proper noun. We can, consequently, disambiguate the meaning of a word to a degree by learning its parts of speech role. Information about each word's parts of speech also allowed us to remove the many words in the scaffolding of language that are unlikely to correspond to concepts. The parts of speech tags provided by GATE were used to narrow our list of words to nouns and adjectives (which qualify nouns) and a handful of words in other categories that seemed pertinent. Furthermore, we narrowed our effort to terms that occur at least 17 times in all the speeches—this is a mere 1% of all the speech paragraphs. No matter how powerfully different the rate of use of the word by the candidates, a term that occurs in only 1% of paragraphs will have low

predictive power.

Identifying nouns and adjectives allows us to identify word "tuples"—that is, combinations of nouns and adjectives that have unique meanings. The tuple "middle-class" can mean something quite different than the words used individually or in other combinations, which might refer to a school class or the ideological middle. Thus, identifying words' parts of speech can be an important step in the direction of identifying particular concepts in people's speech. Fortunately, computer scientists have developed highly accurate methods of identifying a word's part of speech role.

We utilized the open-source General Architecture for Text Engineering (GATE) to process the candidate speeches. In addition to its capacity to tag words with their parts of speech, GATE (Cunningham, Maynard, Bontcheva, & Tablan, 2002) also performs a number of other useful functions, such as clarifying where words and sentences begin and end, identifying punctuation, and so forth. We were able to use the structured data that GATE returns to mark word tuples. GATE is capable of more sophisticated processing that can further disambiguate words, so this paper scratches the surface of what is possible with this technology.

The data was subdivided into a "training" dataset of 56 speeches (approximately 80% of all speeches) and a confirmation dataset of 15 speeches. The confirmation set was determined with a stratified, interval sample. The selection was stratified by speaker. Speeches were ordered by date of speech and then selected at regular intervals. This procedure helps insure a representative split by speaker and by the date of the speech.

Measures

We utilize statistical bootstrapping (Efron & Tibshirani, 1993) as a method of

determining which words are used at significantly different rates between speakers—in our case, the candidates Obama and McCain. Bootstrapping makes no distributional assumptions but rather infers the distribution from resampling of the data itself. Freedom from distributional assumptions is key because the distributions of words may be very complex—affected by contingencies in the speaking context and the context of sentences, paragraphs, and speeches. Indeed, it is unlikely that speakers draw from the same "urn of words" across paragraphs or speeches. The measure we introduce is a bootstrap of the difference between a word's use rates between the two candidates: rate difference=(the word's count for Obama)/(total words used by Obama) – (the word's count for McCain)/(total words used by McCain). If this difference is significantly less than zero, the reverse holds. Define the probability of a rate difference, p(rate dif), as the probability the rate difference is less than zero. If it is highly improbable that the rate difference is less than zero, then it is likely Obama is using the word at a higher rate than McCain. If it is highly probable that the rate difference is less than zero, then it is likely McCain is using the word at a higher rate. The focus, then, will be on values for this probability that are extreme, such as below .001 for Obama and above .999 for McCain. Because we will be looking through over 700 words, a Bonferroni correction might be in order for the p-values. This, however, overcorrects for simultaneous testing, and a p-value of .999 should still draw some attention, particularly if the word identified may in theory have a relationship to differences among ideologies. For this reason, we identify promising discriminating words in a training dataset and then confirm that these words predict to other data, as we do here.

In addition to the issue of whether a given word is used at significantly different rates by different types of speakers, there remains a second issue of whether a given word will have enough impact on predictive capacity to be worth including in a more sophisticated statistical

model. For example, a word that is used at only mildly different rates by the candidates but which occurs frequently in conversation can have a larger effect on predictive power than a word that discriminates powerfully but occurs very rarely. The frequency and discriminative power of concepts both matter to the mental models of the speakers. A concept a speaker uses frequently is more central to the thinking of the speaker. We introduce a measure of predictive power of a word by asking how much knowing that it occurs once in a passage shifts prior probabilities. Specifically, the measure assumes that the prior probability of Obama or McCain being the speaker in a passage is .5 and determines how much change there is in this probability knowing that the speaker has used a single word with a known ratio of occurrence between the speakers. This is easily calculated applying Bayes's theorem. Suppose that the use of the word shifts the probability that the speaker is McCain to .8 from .5. Thus, the net change is .3. Suppose also that the word occurs 10 times overall for both speakers. Then the net change is 10*.3 or 3. The shift, however, will be in the wrong direction in two of those 10 instances, so a full accounting of correct change would subtract these two instances from the eight correct ones. The final total impact is (8*.3 - 2*.3)=1.8. The measure indicates, admittedly with the help of some simplifying assumptions, net correct shift in p-values over all occurrences of a word. One interpretation of the measure is that twice its value is roughly how many passages go from a 50-50 split between candidates to certainty that it is one of the candidates. So, a value of 10 would indicate that the measure, loosely speaking, brings certainty to 20 passages.

The bootstrapped rate difference and the predictive power indicator will help narrow the field of possible words to those that both are likely to really reflect differences in word use and have a larger impact on differentiating the candidates. Ultimately, however, individual words do not occur with enough frequency to appreciably predict which candidate is speaking across all

passages. For real power in predicting who is speaking, it is necessary to combine the impact of a multitude of words in a statistical model that predicts which candidate is speaking. Such a model also helps clarify the mental models the candidates are using because words that in isolation seem predictive may not prove predictive after controlling for other words. The model can also be applied to other datasets, for example to track the degree to which presidential candidate's mental models are reflected in the media. The statistical distribution chosen for this analysis is the Dirichlet multinomial (DM) distribution, which was estimated by maximum likelihood using the R package VGAM. The distribution is a multivariate version of the beta-binomial model, which in turn is a more sophisticated version of the binomial model. The binomial model could be used to model the number of instances of one given word in a passage based on the assumption that the process creating the passage samples that word with a given frequency—in essence, it assumes people draw words as if out of an urn with two colored marbles, one for the word and one for other words. The shortcoming of this model is that text clearly varies in the frequency of words across passages and speeches. The beta-binomial model addresses this deficit by proposing that people draw out of a distribution of urns, each with its own proportion of words. The Dirichlet multinomial model allows for any number of words to be analyzed, rather than just two alternatives, and like the beta-binomial assumes people are drawing from a distribution of urns with different mixtures of words. It therefore more realistically captures real speech.

Results

Data Characteristics

The full dataset contains 1716 observations, each providing data on one paragraph from the two candidates. The data consist of columns each of which contains the counts of a given word. As described in the methods section, only a subset of types of words, such as nouns, adjectives, pronouns, and indicators of reasoning, were selected. Of the 1716 paragraphs, 42% were from Obama and the rest from McCain. The data include 55 speeches, 33% of which were made by Obama. A more balanced dataset should strengthen the results reported here. A total of 134,724 words were examined, 36% of which came from Obama. There were 751 unique words under consideration. Words involved in different parts of speech, such as the adjectival or noun versions of a word, are treated as separate words. The median count for a given word is 38 and the 90th percentile was 158. Thus, even at the 90th percentile, a given word constitutes only .1% of all words. The most prevalent word, "we," constitutes merely 1.9% of all words. As the prevalence of a word increases, its power to differentiate between Obama and McCain in any given instance declines because the proportion of word use for a candidate moves towards 50%. This means that it will take a large number of words in the final model to adequately predict who is speaking.

For purposes of analysis, the full dataset was subdivided into a "training" and a "confirmation" set, with most of the analysis reported below coming from the training dataset. Fifteen speeches were set aside, using stratification by candidate and selecting by interval ordered by date of speech, for the confirmation set and the rest were used in the training set. General characteristics of the training and confirmation sets were similar.

Selecting Words

The first step in word selection involves examining the probability that the difference in word use rates is negative. A p-value near 0 indicates that the rate difference is most likely positive and therefore Obama uses the word at a higher rate than McCain. A p-value near 1 indicates the opposite. P-values were determined by bootstrapping speeches. One simple interpretation of the result is that it indicates how robust the rate difference is across speeches. Another interpretation is that it implies that the mechanism that selects words for speeches, whatever it is, involves a rate difference between the candidates.

The first point that needs to be established is that the p-values of the differences in rates of word use by Obama and McCain are indeed highly improbable, despite the large number of words examined. Figure 1 shows a graph that plots the bootstrapped p-values, one for each of the 751 unique words, against the quantiles of the words. If the p-values occur simply by chance, the plot of the p-values should closely follow the 45 degree line on the graph. In fact, however, there are far more p-values near 1 and near 0 than would be expected by chance alone. Another interesting implication of Figure 1 is that there are appreciably more words whose use indicates McCain is speaking than words indicating Obama is speaking. This may have broader implications about ideology, as will be examined in the discussion section.

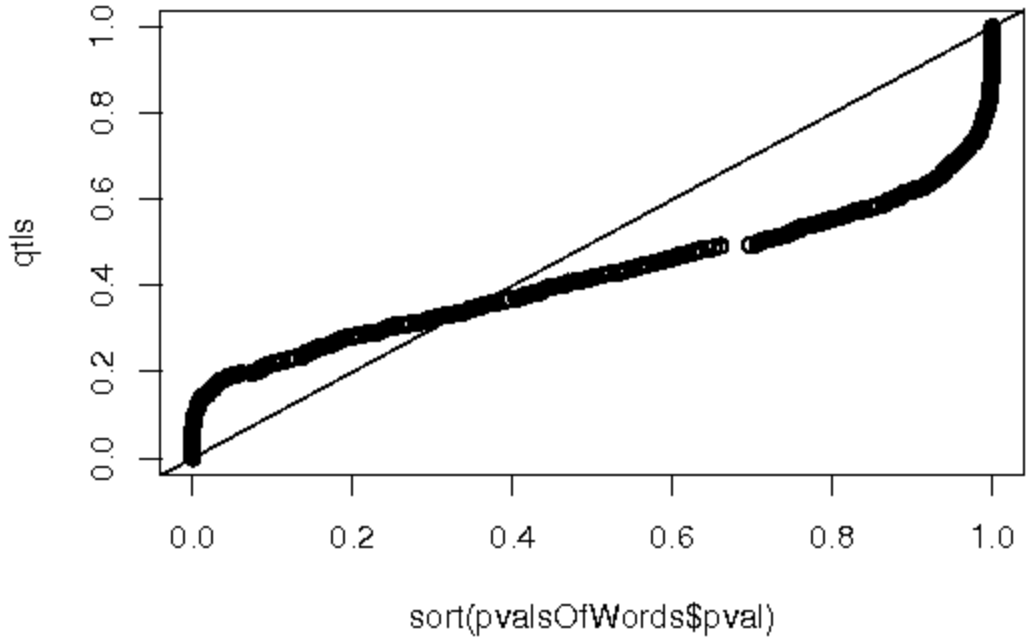
qtls
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
sort(pvalsOfWords$pval)

Table 1 depicts selection indicators for a subset of the words. p(rate dif.) is the p-value for the hypothesis that McCain uses the word at a higher rate than Obama. Values near one suggest that McCain in fact uses the word at a higher rate than Obama. Values near zero indicate that the word is used more frequently by Obama. The Predictive Power indicator, as discussed earlier, shows how much net correct movement there would be in the *a posteriori* probability that paragraphs are from one of the candidate or the other were the observer only to know which paragraphs contained an instance of the word and the proportion of the words used by each candidate. p(Obama) indicates the proportion of all instances of the word's use that actually came from Obama, and N of Word is the total number of instances of the word in all speeches in the "training" dataset, which represents about 75% of all the speeches. Table 1 shows many of the words that made the cut for inclusion in the final Dirichlet multinomial model. The cut was determined by taking the top 50 words in terms of Predictive Power that were also highly significant for p(rate difference), either near zero or near 1. Generally, p(rate difference) was above .99 or below .01, though an occasional exception was made for interesting and theoretically important terms. Table 1 presents 37 of the words from the top 50 that could be readily classified into the categories of the table. A more complete analysis would look at more than the top 50 words, but we are interested here in testing the waters.

Table 1 groups words into interpretive categories. These categories were determined by a non-systematic examination of the words in their speech contexts and are offered here as proof of concept. A more rigorous analysis involving human content coding will be important to helping establish the categories and classifications suggested here. Also, comparison with the speeches of others, such as vice presidential candidates or candidates in historic races, may help clarify what actually are personal stylistic differences, presentational choices, or differences that

are important to certain ideological and political orientations. For now, we will offer some provisional interpretations.

The category of policy issues concerns specific references to concepts concerning public policy options. For example, McCain typically discusses "market(s)" in contexts having to do with policies favoring free markets. He mentions government with respect to policies to curtail and restrain government spending. He discusses unemployment and health insurance, particularly portable health insurance. And, he discusses nuclear energy, nuclear plants, and so forth with respect to energy security and addressing global warming. Interestingly, he mentions "global warming" far more frequently than Obama, as part of discussions of environmental stewardship. In contrast to McCain, there are few cases of Obama using words related to public policies at a higher rate than McCain.

As with policy issues, McCain deploys more words describing targets of policy that differentiate him from Obama. As might be expected from typical liberal concerns, Obama discusses children and higher education (colleges) more often than McCain. Perhaps also consistent is his use of the notion of community, which refers to concrete communities. This contrasts with McCain's more frequent use of the word "public" as an abstraction—*the* public or the public / private distinction. Obama has a higher rate of using "family," often in the context of discussing economic hardships. McCain does not offset this reference to families with discussion of the conservative issue of family values, which he never mentions. In fact, it is only Obama who discusses family values in the form of criticisms of conservative politicians who stress family values but whose policies do not promote the economic well-being of families.

Two organizing concepts might help explain much of McCain's word use: republicanism and free market ideology. Some variants of republicanism value the state,

institutions, and authorities over individuals. People are seen as subordinate to something bigger than themselves, which is embodied in the state, institutions, history, and greatness. In addition to republicanism, McCain also stresses words that are associated with free market ideology. This ideology might represent a tension in McCain's thought, because it could be in opposition to republicanism. Alternatively, it might simply be lip service to an important constituency, or perhaps it is integrated into a republican belief system—for example, by valuing free markets as useful to the state. A Bayesian network analysis would be helpful in determining which of these is the case.

Free market ideology may explain McCain's references to markets, portable insurance (which would liberate employers from having to pay for insurance), and small business. Mention of capable persons and leaders is consistent with free market ideology and Republicanism. Republicanism explains many of the other references. Republicanism can involve a concern with environmental stewardship, for the long-term good of the country. This makes sense of McCain's concern with nuclear energy to protect the planet from global warming, another term McCain uses more than Obama. Republican focus on the value and character of institutions is consistent with McCain's disproportionate mention of institutions—the courts, Congress, federal government and federal policies, and international institutions and relations. Concern with or vaunting of institutional authorities is consistent with his more intensive use of terms such as "judges," "officer," and "judicial." It is also consistent with the focus on the abstract public and the public / private distinction, rather than on concrete communities as in the case of Obama. Republican attention to the greatness and importance of a collective that transcends the concerns of the individual help explain McCain's more intensive use of such terms as "power," "great," "important," "serious," "purpose," and perhaps "human" in such forms as

"human history" and "human beings." Republican thought patterns may also clarify why McCain stresses the other: other peoples, other countries, and so forth. Republicans define the self in terms of collectives, hence people in different collectives constitute distinctive others to one another. This may be in some tension, however, with the overall notion of "human beings" McCain at times uses. Finally, McCain does not focus on social divisions, as Obama does with his significantly greater use of "white" and "black." This is consistent with the republican notion of the unity of the nation and the public.

In contrast, the words that Obama uses suggest a focus that is more local, personal, egalitarian, and inclusive. He speaks of family, children and community more than large institutional actors. His policy issue focus is on something with which people have everyday interactions—schools—rather than such abstractions as energy security and markets. Perhaps Obama's community organizing work influenced this preference for the concrete. When it comes to words indicating explanation, Obama uses logical connectors such as "if" and "because" and signifiers of explanation such as "why" significantly more often than McCain. Explaining matters to the audience implies greater equality between the speaker and the audience—the audience is viewed as having a right to an explanation. In contrast, republicans along the lines of George Will's (1983) take the view that the public is subordinate to leaders of state and imply the public should not demand explanations of the goals and purposes of the state. Obama's more prevalent use of the term "we" might also imply a greater willingness to put himself at an equal level with his audience. A human content analysis of the use of the term may prove revealing. Stylistically, Obama uses contractions such as "she'll" and "he'd" far more than does McCain. Thus, he presents himself less formally and, therefore, at a more equal level to his audience. The the extent that Obama speaks of a larger collective purpose, it is in terms

of asking people to consider themselves part of a larger historical story—hence his more prevalent use of "story." In place, then, of a transcendent and demanding collective, Obama offers the metaphor of a narrative into which people might write their own lives and experiences.

Table 1: Some Words that are Significantly Different Between the Candidates

Word	p(rate dif.)	Predictive Power	p(Obama)	N of Word
Policy Issues				
energy security	1.00	12.0	.00	24
market	1.00	42.3	.08	122
nuclear	1.00	50.5	.11	162
insurance	1.00	12.1	.12	42
government	1.00	32.1	.27	296
schools	.00	14.9	.86	60
Policy Targets				
court	1.00	14.9	.05	37
judges	1.00	14.4	.05	36
small business	1.00	20.7	.09	60
Congress	1.00	22.5	.10	70
international	.99	18.7	.14	72
federal	1.00	29.0	.19	154
public	1.00	13.2	.26	112
family	.00	12.7	.71	140
children	.00	23.1	.79	143
community	.00	20.8	.85	86
college	.00	16.5	.98	36
Social and Interpersonal Divisions				
capable	1.00	17.0	.00	34
other	1.00	20.3	.34	402
black	.02	31.8	.97	72
white	.00	35.0	1.00	70
Orientation Toward Power, Authority, Government				
judicial	.99	28.0	.00	56
officer	1.00	13.5	.00	27
human	1.00	17.0	.08	46
purpose	1.00	12.6	.09	37
serious	1.00	28.0	.11	92
power	1.00	21.3	.23	147
important	1.00	19.0	.24	136
great	1.00	31.1	.28	324
we	.01	15.3	.56	1933
'll	.00	53.3	.91	161
story	.00	20.5	.91	60
Explanatory Style				
if	.00	13.1	.68	210
because	.00	20.4	.71	241
why	.00	22.9	.85	93
Political Strategy				
Canada	1.00	17.5	.00	35
George Bush	.00	13.0	1.00	26

Multivariate Modeling

Two Dirichlet multinomial models were estimated using the 50 words that were selected as most valuable for discriminating between the candidates. The first model was for Obama's paragraphs and the second was for McCain's. Each model estimates parameters that, when applied to the word count variables, best describes the distribution of words for the model's candidates. Most of the output consists of hyperparameter estimates that are not readily interpreted, so we do not give details of the results here. The one parameter that is readily meaningful is phi, the dispersion factor estimate. Values near infinity would indicate that the model is not overdispersed—that is, each paragraph is drawn from the same probability urn of words, not from a distribution of urns. That parameter's values for the Obama and McCain models, respectively, were .05 and .02. Paragraphs are not drawn from the same probability urn.

The two models for candidate's word distributions can then be applied to a data set, either the original training set on which the models were developed, or the confirmation set. A minor Bayesian manipulation of the probability densities from the models converts the densities into the probability that the paragraph was uttered by Obama rather than McCain (or vice versa). The value of these probability estimates can then compared to prior knowledge. Suppose that people making predictions of who spoke which paragraph in the training dataset know that 58% of the paragraphs are from McCain. Then they could maximize their predictive power by predicting that all the paragraphs are from McCain. They would be right 58% of the time. With the probability estimates from the models, they would be right 78% of the time—a 34% increase in the number of correct predictions (from the base of 58%). A probit analysis regressing an indicator of which candidate uttered each paragraph on the probability estimates

for each paragraph indicates the latter significantly predicts the former with a p-value far in excess of .0001.

Applying this procedure to the confirmation dataset, we find that with prior information observers could correctly predict 55% of the paragraphs. With the probability estimates from the model, the observers would be correct 69% of the time, a 25% increase in the number of correct values. A probit regression indicates the probabilities significantly predict who is speaking, also with a p-value far in excess of .0001.

The predictive power of the model needs to be considered in light of the fact that the median paragraph has 88 or fewer words—about two and a half times the length of this sentence. Larger units of text, such as entire speeches, would have a far higher rate of correct prediction. Indeed, applying the model to predicting who gave which speech in the confirmation (non-training) data increases the accuracy of prediction to 93%, from 60% correct in the prior.

Discussion and Conclusion

This paper examines whether a moderately simple set of statistical tools and application of natural language processing can help extract meaning from text, particularly by helping to reveal the "mental models" that inform the authors of the texts in their choices of concepts to discuss. We examined speeches by John McCain and Barack Obama through July 9, 2008. The purpose of the paper is to present a proof of concept for the tools and the basic methods deployed. Subsequent revisions will examine all candidate speeches and add more sophisticated methods. The core assumption made in this early version of our methodology is that speakers with different political leanings will use somewhat different sets of concepts, that these

differences can be captured by examining which words speakers use and will successfully identify important aspects of differences between speakers, and, in particular, that the differences in word use between speakers will reveal interesting differences in speaker's mental models.

Figure 1 shows that indeed there are strong differences between McCain and Obama in their choice of words. To help demonstrate that the words identified as significantly different between McCain and Obama capture important differences between them, we present results of a multivariate model that finds that the words powerfully predict which speaker uttered a given paragraph or speech, even for data on which the models were not built. This model was limited to only 50 words out of a body of more than 6,051 unique words used by both speakers. In future work, we will consider more complete models selected with Holm-Bonferroni p-values. One interesting test for future work would be to determine how well the model fares relative to human judges of who is speaking.

To establish that word use differences between speakers help reveal the speaker's mental models, we offer, as a preliminary step, a non-systematic interpretation of many of the 50 most discriminative words, listed in Table 1. We find the words used suggest deep and interesting ideological differences between the speakers. John McCain appears to embrace a conservative form of the political philosophy of republicanism, with some references pertinent to free market ideology. This form of republicanism exalts the purpose and value of the state rather than the individual. Thus, we find that McCain disproportionately uses words about greatness, seriousness, and purpose. He speaks disproportionately about institutional authorities, while Obama references the personal and local: family, children, and community. The policy issues McCain disproportionately mentions are abstract and concern the long-term welfare of the country. Obama, in contrast, stresses something with which people have day-to-day

experiences—schools. Unlike Obama, McCain does not focus on typical social divisions ("white", "black"), but does stress otherness (other peoples, other countries) and distinguishes some people as "capable." Obama presents himself informally, by using contractions at much higher rates than McCain. His use of explanatory words ("if," "because," "why") likewise suggest greater equality between himself and the audience. Obama, then, presents himself as a populist: informal, egalitarian, and concerned with issues on a small, personal scale: community, family, and so forth. These interpretations are interestingly contrary to the perception of many political commentators that Obama conveyed an elitist image while McCain did not. To establish these interpretations, we will pursue a systematic content analysis of random sentences containing the key words. Such an analysis will take far less time than a full content analysis of the speeches.

One intriguing finding is that there are far more words that McCain uses significantly more often than the reverse. This is consistent with an earlier unpublished analysis of conservative and liberal speech on online discussion boards for students taking a required American public policy course. Here as well, there were appreciably more words indicating conservative speech than liberal speech. Perhaps this is the product of the concerted effort by the conservative movement to create conservative think tanks and spokespersons for their movement and to coordinate publicity efforts. Conservatives appear to have adopted a more unique vocabulary than liberals, which may indicate a more unique set of concerns or the use of verbal signals to indicate preferences.

An important question is how much of people's mental models can ultimately be discerned by a focus on differences in the rate of word use between speakers. Certainly there are important differences that may not be captured by identifying words with strong rate

differences. For instance, both McCain and Obama speak about as frequently as the other about Iraq, yet their policy positions regarding continuing that war are quite different. A response to this criticism is that while the methods described here may not be able to discern every difference between speakers, the methods nevertheless do a good job of predicting who is speaking, suggesting that they are capturing many of the important differences. How well the methods do should perhaps be compared with human judges. Another response to the criticism is that nothing prevents these methods from being used on data that captures the direction of the speaker's sentiment with regard to a word. If that sentiment can be labeled as positive or negative, then a word with a negative valence can be treated by the methods here simply as a different concept than the same word with a positive valence. A final response is that the methods defined here are only a first step and additional processing of the data may help get at some of the subtleties of meaning. The methods here help winnow a dataset containing over 6000 unique variables (words) to only 50. Reducing the vast number of linguistic possibilities to a manageable number is crucial. The next step will be to discern the network of conditional relationships that exist among the words. This analysis can begin with the words that are known to be significantly different between the speakers, but also include words with significant conditional relationships that are not used at significantly different rates between speakers. The resulting pattern of relationships may reveal valence, not simply rate differences. While McCain and Obama do not differ with respect to the word "Iraq," they do differ on words such as "terrorists," which are likely to have different conditional relationships with words such as "Iraq." Thus, a conditional network analysis could reveal differing sentiment on Iraq. The hope is that most deep conceptual differences between speakers will likely be reflected in at least some words being used differentially between the speakers. Once these are identified, it is

possible to search the network of connected words to bring the full differences into view. The advantage of starting from the words that are significantly different is that starting with all words would require unattainable amounts of computing power.

More work also needs to be done to insure that differences captured by the approach described in this paper are indeed due to real differences in mental models rather than simply strategic choices or personal styles—though perhaps to an extent, strategic choices and personal styles are relevant as well. Nevertheless, to sort out these differences, comparisons between the same speaker in different contexts or between different speakers that are on the same side of an issue would be helpful. The candidates' comments in the debates are less scripted and may reveal more than stump speeches about their underlying positions. Likewise, comparing a candidate with another similar conservative politician's speeches might reveal what is central to an ideology and what is not.

References

Cardie, C., & Wilkerson, J. (2008). Guest editors' introduction: Text annotation for political science research. *Journal of Information Technology & Politics, 5*(1), 1-6.

Coffey, D. (2005). Measuring gubernatorial ideology: A content analysis of State of the State speeches. *State Politics and Policy Quarterly, 5*(1), 88-103.

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). Gate: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Annual Meeting of the ACL*. Retrieved November 17, 2008, from http://citeseer.ist.psu.edu/context/2035358/0.

Dyson, S. B. (2008). Text annotation and the cognitive architecture of political leaders: British Prime Ministers from 1945-2008. *Journal of Information Technology & Politics, 5*(1), 7-18.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.

Jurasfsky, D., & Martin, J.H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Upper Saddle, NJ: Pearson/Prentice Hall.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review, 97*(2), 311-331.

Manning, C.D., & Schutze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: The MIT Press.

McBurnie, K., & Goyal, A. (2007). Feature extraction, automatic hierarchy building, and sentiment analysis of political blogs.

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

Riffe, D., Lacy, S., & Fico, F.G. (1998*). Analyzing media messages: Using quantitative content analysis in research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Russell, S., & Norvig, P. (2002). *Artificial intelligence: A modern approach* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Will, G. F. (1983). *Statecraft as soulcraft: What government does.* Simon & Schuster.

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