

# Watching the Words of the Candidates: Our Research Methods

For the last several years, a group of us in academics and the private sector have been studying how the words people use reflect their social and psychological worlds. We have amassed hundreds of thousands of speeches, natural conversations, novels, poems, blogs, interviews, and other language samples. Using state-of-the-art computerized text analysis methods, we are trying to understand some of the subtleties of natural language use among people across a wide array of settings.

For the 2008 election, we are particularly interested in how the major candidates use words in their speeches, press conferences, debates, and interviews. Because we have similar data sets for the 2000 and 2004 elections, we can make some comparisons with past candidates. Readers interested in some of our research on other elections and political leaders can go to our [general research website](#) or click on specific articles listed at the end of this page.

## 2008 Election Debates

Several of the WordWatchers entries in January and February are based on several debates among the leading Democrats and Republicans. The dates, sponsor, and location of the debates that we have used include:

### *Democratic debates*

June 3, 2007	CNN, Manchester, New Hampshire
November 15, 2007	CNN, Las Vegas, Nevada
December 4, 2007	NPR, Des Moines, Iowa
January 5, 2008	ABC/facebook, Manchester, New Hampshire
January 15, 2008	MSNBC, Las Vegas, Nevada

### *Republican debates*

June 5, 2007	CNN, Manchester, New Hampshire
November 28, 2007	CNN, St. Petersburg, Florida
December 12, 2007	PBS, Johnstown, Iowa
January 5, 2008	ABC/facebook, Manchester, New Hampshire
January 10, 2008	FOX, Myrtle Beach, South Carolina

Transcripts of the debates are available through ProCon.org. The transcripts of each debate were downloaded and cleaned of transcriber comments (e.g., notes such as “laughter”, “inaudible”, “applause”, “commercial break” were removed). All words used by the major candidates were then separated and put into individual text files, one for each debate. For the first five Democratic debates, then, five separate files were created for Clinton, Edwards, and Obama. For the Republicans, five files were made for each of the major contenders at the time: Giuliani, Huckabee, McCain, and Romney.

Across the five debates, each of the seven major candidates were impressively talkative

ranging from 1,924 words per debate (McCain) to 3,742 (Clinton). On average, the Republican candidates spoke fewer words per person (2,270) than did the Democrats (3,389) due to the fact that there were always more Republican debaters than Democratic.

## Text and Data Analysis Methods

Unless noted otherwise, most of the text analysis methods relied on a computer program called LIWC or Linguistic Inquiry and Word Count (Pennebaker, Booth, & Francis, 2007). LIWC, which is a commercially available software program ([www.LIWC.net](http://www.LIWC.net)), analyzes text samples on a word-by-word basis, and compares each to a dictionary of over 2,000 words divided into 74 linguistic categories.

Output is expressed as a percentage of the total words in the text sample. For example, use of 1<sup>st</sup> person plural pronouns (e.g., we, us, our) varies significantly by candidate. Through LIWC, we have calculated that Edwards uses we-words at 2.7 percent of all the words he uses. On the other hand, 4.2 percent of all of Romney's words are 1<sup>st</sup> person plural.

It should be noted that some of the categories are defined purely grammatically. For example, the "articles" category searches for instances of a, an, and the. Other categories, such as positive emotion words, were formed initially by having independent judges decide which words should go into each category. One of the strengths of a word counting program such as LIWC is that it does an excellent job in capturing common words that people use in daily conversation. For the debate texts, LIWC recognizes about 88% of the words that any candidate speaks—proper nouns and very low-frequency words comprise the other 12%.

For the debates by candidate, we use a straightforward analysis of variance (ANOVA) strategy using candidate as the independent variable and each language dimension within each debate as the unit of analysis. So, for example, the analysis of 1<sup>st</sup> person plural yields the following information for the candidates:

<b>Candidate</b>	<b>Number of Cases</b>	<b>Mean</b>	<b>Standard Deviation</b>
Clinton	5	3.40	0.43
Edwards	5	2.68	0.57
Obama	5	4.02	0.51
Giuliani	5	3.16	0.96
Huckabee	5	3.67	0.93
McCain	5	4.14	0.78
Romney	5	4.16	0.80

A simple one-way ANOVA indicates that the use of we-words differ across the various candidates,  $F(6,28) = 2.84, p = .027$ . Note that this analytic strategy is conservative (in that we are falsely assuming a between-subjects strategy with independent assumptions) and based

on only five observations – one for each debate. In addition, we will be focusing on approximately a dozen language dimensions.

## **Disclaimer and Caveats**

Should any researchers like to receive a copy of the data, please contact us ([Pennebaker@mail.utexas.edu](mailto:Pennebaker@mail.utexas.edu)). The information that we are providing should be viewed with caution. Our goal is to give readers a glimpse of some of the developing trends in the word use of the candidates. Our interpretations are preliminary. Any ideas or suggestions about the data, the analyses, or interpretations would be greatly appreciated.

James W. Pennebaker  
Professor of Psychology  
University of Texas at Austin