Saying Whatever It Takes: Creating and Analyzing Corpora from US Presidential Debate Transcripts

Leo Vrana (University of Konstanz, Germany) and Gerold Schneider (University of Konstanz, Germany/University of Zurich, Switzerland)

We first describe the creation of a corpus of American presidential debates from the *American presidency project*. We then use the corpus to present a stylistic analysis of presidential candidates from 2000 to 2016. A range of stylistic measures, including vocabulary richness, language complexity, and readability measures is applied. We aim to contribute to the current debate on the complexity of American presidential rhetoric and the role of the register of spoken language, by furnishing empirical data[1].

1. Introduction

In the American news media, United States Presidential elections are dissected and examined with a fervor normally reserved for sports championships and natural disasters. This is not surprising given what is at stake. Our contribution focuses on the creation of a corpus of Democratic and Republican presidential nominees' speech during these debates dating back to the year 2000, and presents stylistic results gleaned from this corpus, including measures of vocabulary richness, language complexity and readability. Further, we compare these results against another corpus of spoken American English in order to possibly uncover any further insights into the differences between the speech of politicians and everyday citizens.

While word-choice and topics vary depending on the agenda of the candidate and on current issues, stylistic features can be varied freely to convey a message tailored to targeted voters. Lim (2008) has claimed that American presidential rhetoric has become considerably less complex over time, a trend that created lively discussions during the election of George W. Bush and even more so with Trump's election. Simpler language in speeches may signal an attempt to address a broader, less educated audience and to satisfy the demands of "sound bite" journalism (Hallin, 1992), but may also be a way to address the demands of the spoken genre, where high complexity increases the risk of ambiguity. This paper also aims to provide empirical data to complement these analyses.

2. The Source and Corpus Creation

The American Presidency Project is a non-profit undertaking hosted online by the University of California at Santa Barbara which archives documents related to United States presidents for public use, including transcriptions of the presidential debates, accessible at http://www.presidency.ucsb.edu/. The first goal of this project was to

[1] Files related to this project can be found at *https://github.com/LeoVrana/PresidentialDebates*.

take these transcriptions and compile utterances of each candidate into a corpus for that candidate.

In order to accomplish this, we took the page source for each transcript, and used regular expressions and capturing groups to automatically create a corpus for each candidate. We included a manual validation step to ensure that the data was processed correctly.

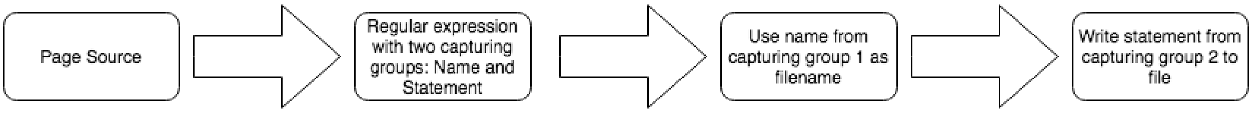

Page Source
Regular expression with two capturing groups: Name and Statement
Use name from capturing group 1 as filename
Write statement from capturing group 2 to file

Fig. 1: Flow Chart

3. Methods and Results

After ensuring that the files were accurate, we performed the analyses below. We also include figures from the Santa Barbara Corpus of Spoken English (SBC) as a reference where possible (Dubois, Chafe, Meyer & Thompson, 2000-2005). Statistical significance tests were performed to compare politicians relative to each other – values from the SBC were not considered.

3.1 Word-Level Analyses

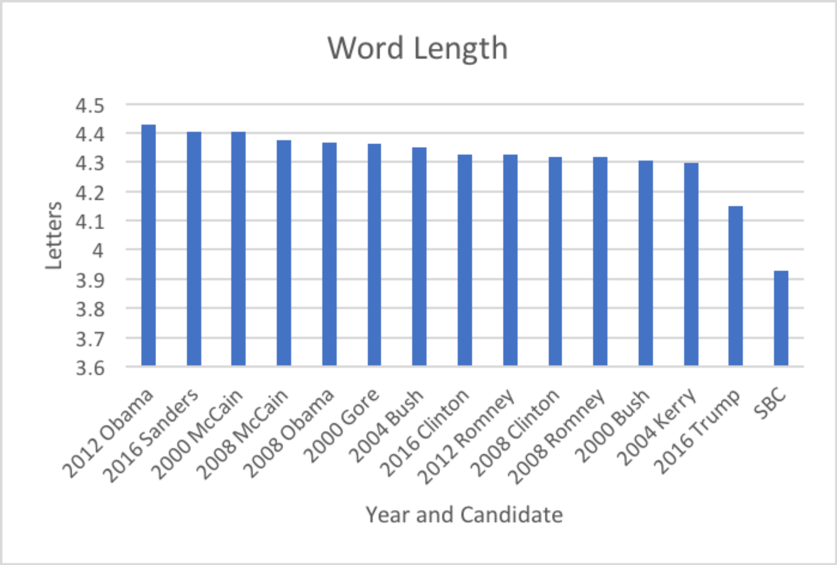


Fig. 2: Mean of word length in letters

The mean word-length shows that candidates were not far apart from each other, with the statistically significant exception of Donald Trump (z = -2.89, p < 0.01). The average word length from Santa Barbara Corpus was similarly shorter than the other politicians, perhaps partly due to the transcription method. The related count of average syllables per word shows very similar results. Trump was found significantly differ from other candidates here as well (z = -2.8, p < 0.01).

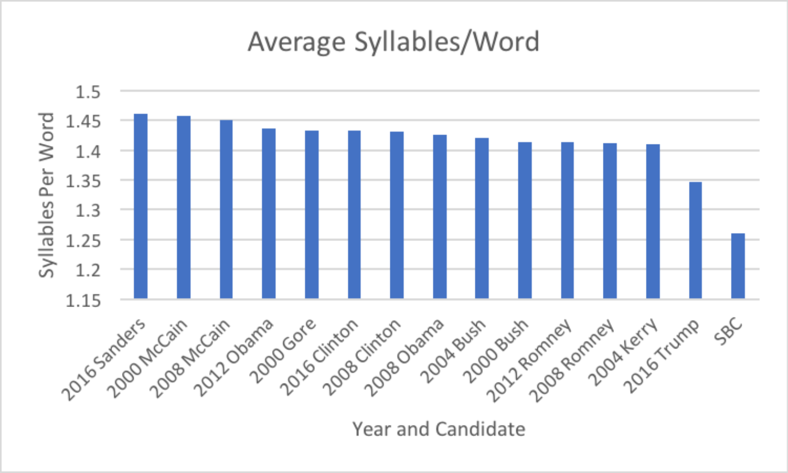


Fig. 3: Mean of syllables per word

3.2 Vocabulary Richness

Measures of vocabulary richness are typically based on type-token ratios. Since tests of vocabulary richness are affected by the size of the corpus, any analysis must account for this (Malvern et al. 2004, Lu 2014: 82). We used the Mean-Segmental TTR (MSTTR), which calculates TTRs for segments of text, and then finds the mean of those TTRs (Lu 2014: 82). A smaller TTR indicates less varied speech.

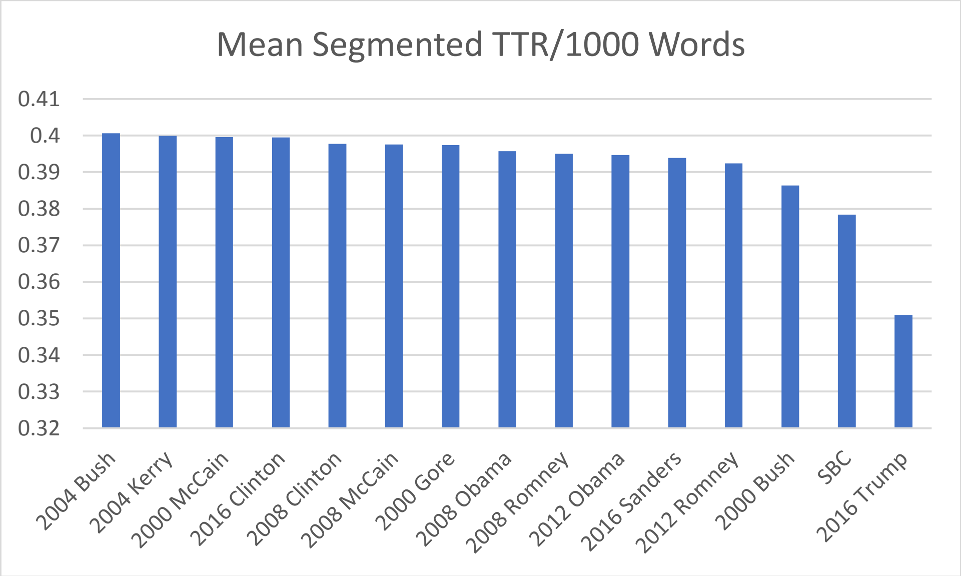


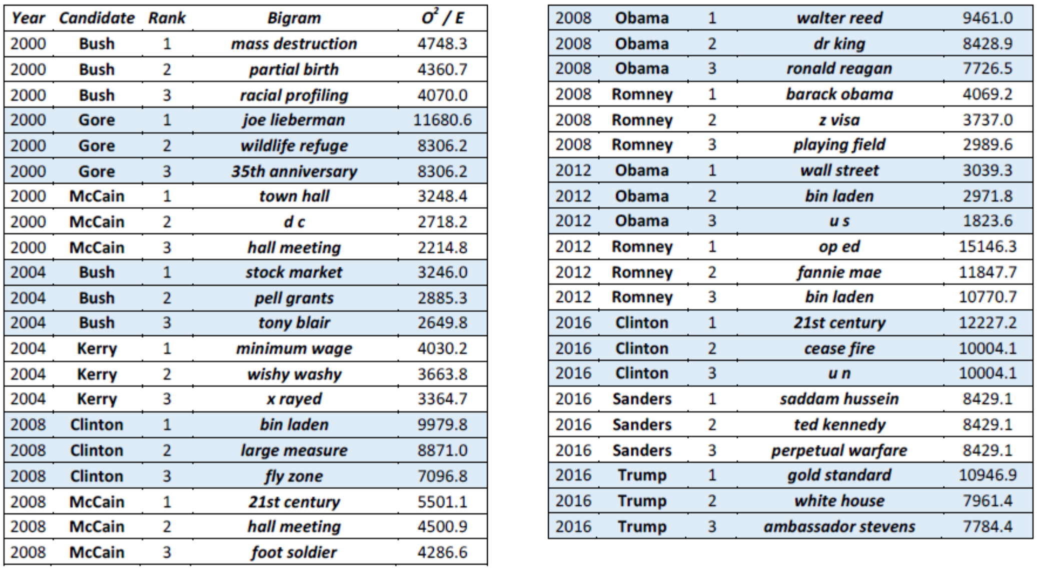
Fig. 4: Mean-Segmented TTR

The measurement shows a large variance, with Trump's vocabulary being significantly less varied than other politicians ($z = -2.84$, $p < 0.01$), as well as the SBC.

3.3 Bigram Collocations

Analyses of collocations, in addition to linguistic multi-word entities, often bring up key concepts, particularly when analyzing monotopical texts. Collocation research has a long tradition, see Pecina (2009) or Evert (2009). We use the measures Observed over Expected (O/E) and its variant O^2/E, which are simple to calculate and interpret, and have a tendency to over-report collocations consisting of rare, i.e. content, words, thus giving insights into candidates' agendas and core interests.

We list the top three bigrams for each candidate below along with the campaign year, sorted by O^2/E (last column).



Year	*Candidate*	*Rank*	*Bigram*	*O^2 / E*
2000	**Bush**	1	***mass destruction***	4748.3
2000	**Bush**	2	***partial birth***	4360.7
2000	**Bush**	3	***racial profiling***	4070.0
2000	**Gore**	1	***joe lieberman***	11680.6
2000	**Gore**	2	***wildlife refuge***	8306.2
2000	**Gore**	3	***35th anniversary***	8306.2
2000	**McCain**	1	***town hall***	3248.4
2000	**McCain**	2	***d c***	2718.2
2000	**McCain**	3	***hall meeting***	2214.8
2004	**Bush**	1	***stock market***	3246.0
2004	**Bush**	2	***pell grants***	2885.3
2004	**Bush**	3	***tony blair***	2649.8
2004	**Kerry**	1	***minimum wage***	4030.2
2004	**Kerry**	2	***wishy washy***	3663.8
2004	**Kerry**	3	***x rayed***	3364.7
2008	**Clinton**	1	***bin laden***	9979.8
2008	**Clinton**	2	***large measure***	8871.0
2008	**Clinton**	3	***fly zone***	7096.8
2008	**McCain**	1	***21st century***	5501.1
2008	**McCain**	2	***hall meeting***	4500.9
2008	**McCain**	3	***foot soldier***	4286.6
2008	**Obama**	1	***walter reed***	9461.0
2008	**Obama**	2	***dr king***	8428.9
2008	**Obama**	3	***ronald reagan***	7726.5
2008	**Romney**	1	***barack obama***	4069.2
2008	**Romney**	2	***z visa***	3737.0
2008	**Romney**	3	***playing field***	2989.6
2012	**Obama**	1	***wall street***	3039.3
2012	**Obama**	2	***bin laden***	2971.8
2012	**Obama**	3	***u s***	1823.6
2012	**Romney**	1	***op ed***	15146.3
2012	**Romney**	2	***fannie mae***	11847.7
2012	**Romney**	3	***bin laden***	10770.7
2016	**Clinton**	1	***21st century***	12227.2
2016	**Clinton**	2	***cease fire***	10004.1
2016	**Clinton**	3	***u n***	10004.1
2016	**Sanders**	1	***saddam hussein***	8429.1
2016	**Sanders**	2	***ted kennedy***	8429.1
2016	**Sanders**	3	***perpetual warfare***	8429.1
2016	**Trump**	1	***gold standard***	10946.9
2016	**Trump**	2	***white house***	7961.4
2016	**Trump**	3	***ambassador stevens***	7784.4

Table 1: Top 3 collocations per campaign

3.4 Readability

Readability measures partly depend on sentence length. Fortunately, the debate transcriptions added punctuation which made this analysis possible. We determine the ease to which a text can be understood using the *Lingua::EN::Fathom* module by Kim Ryan (available at cpan.org). One of the offered measures is the percentage of "Complex Words," where a word is considered "complex" if it contains three or more syllables.

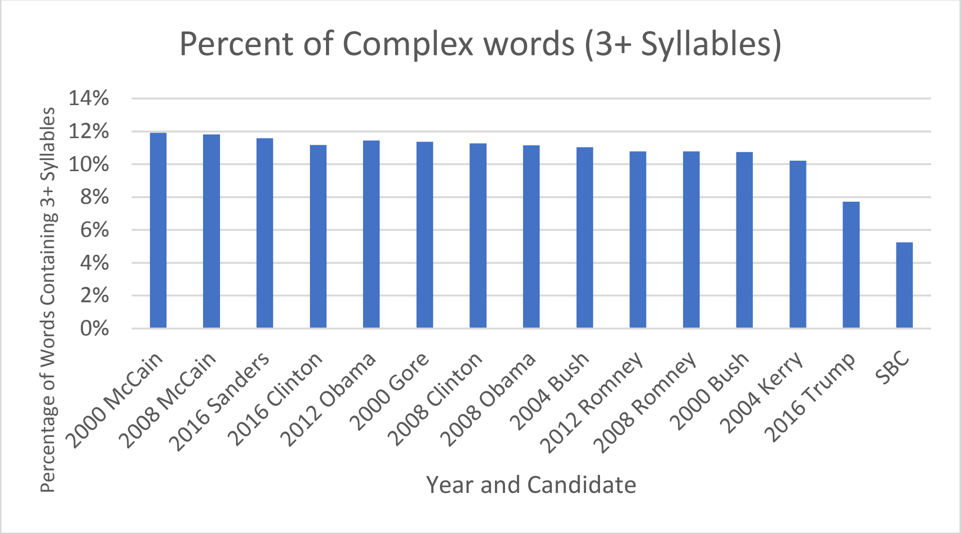


Fig. 5: Percentage of Complex Words

One reason for the radical difference for SBC is that the transcriptions in that corpus include hesitation words such as *"uh"*. Trump was found to differ significantly from other candidates (z = -3.16, p = 0.001). Another measure of readability is words per sentence:

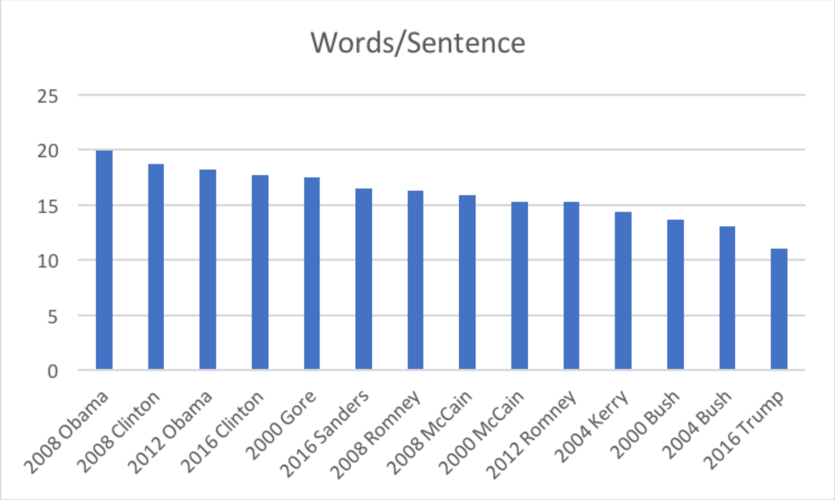


Fig. 6: Words per Sentence

No candidate varied significantly from the rest, but it is interesting to note that with the exception of John Kerry in 2004, Democratic candidates spoke with longer sentences and Republican candidates spoke with shorter sentences. We could not include a comparison to the Santa Barbara Corpus, as the transcription did not include punctuation.

The Flesch readability analysis equation returns a score from 0 to 100, where 100 represents a very easily readable text, and 60-70 is the ideal range (Flesch, 2016). Trump is measured here as being significantly more readable than the other candidates (z = 2.74, p = 0.006).

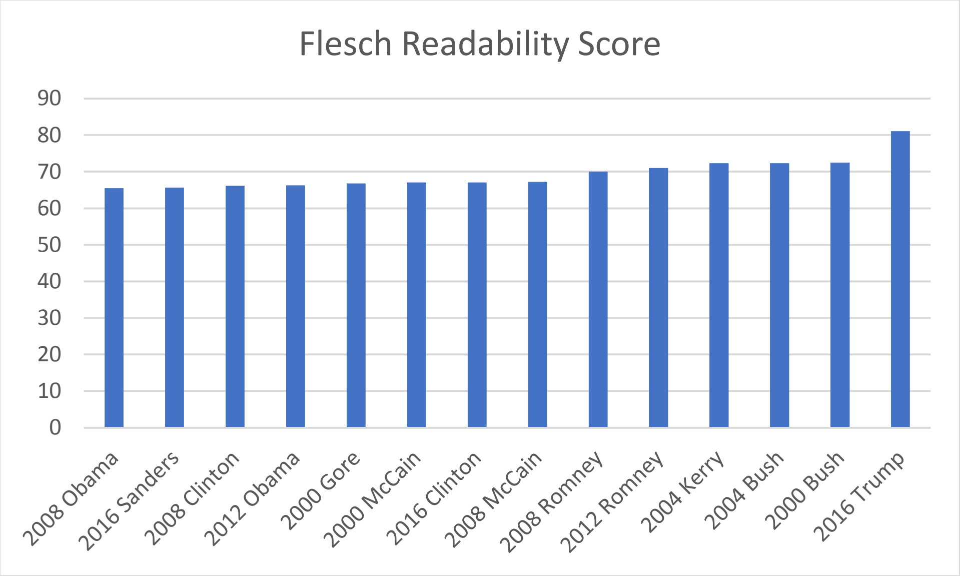


Fig. 7: Flesch Readability Score

While such measures cannot check for grammaticality, they summarize the relative difficulty of understanding each candidate. (Ryan, 2016)

3.5 Surprisal

Readability measures consider words in isolation, but they do not take word sequences into consideration. Psycholinguistic research has shown that routinized sequences are an essential component for ease of readability (Conklin & Schmitt 2012).

We apply surprisal, an information-theoretic measure of the surprise of the continuation of word sequences (Levy and Jaeger 2007). Bigram surprisal is defined as follows:

$$2\text{–} \, gram\ surprisal = log\frac{1}{p(w_1)} + \log\frac{1}{p(w_2|\, w_1)}$$

Surprisal is the logarithmic version of the probability seeing word w_1 linearly combined with the probability of the transition to the next word, w_2. The probability $p(w_1)$, is context-independent, while the transitions, e.g. $p(w_2|\ w_1)$ express predictability in the context. Surprisal is an information theoretic measure; it measures how many bits of information the conversation contains (Shannon 1951): the more expected and thus probable a word is in its context, the less information it carries.

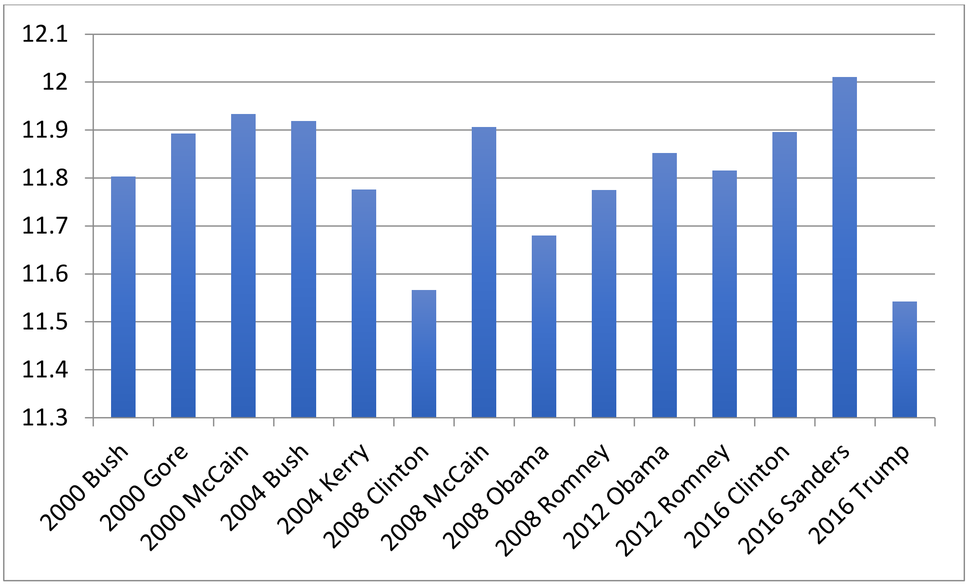


Fig 8. Mean of bigram surprisal

4. Conclusions and Outlook

Our analyses confirmed basic intuitions about speaking patterns of certain candidates, and the concordances and bigrams were interesting reminders of the topics of earlier elections. Differences between transcription methods of the SBC and the presidential debates may be partly responsible for SBC statistics being generally lower than the other politicians, but it must be noted that its statistics were mostly close to Trump's, and both of those were often quite different from other presidential candidates. This may be seen as quantitative evidence supporting the general public's perception of Trump as different from all other candidates (irrespective of the content of his speech, which was outside the scope of this paper), and a candidate who sounded more like a "regular" person.

Opportunities for further research could include sentiment analysis, or analysis of metrics such as audience applause or laughter associated with each candidate.

References

Chissom, Brad S, Fishburne, Robert, Kincaid, J. Peter, and Richard L. Rogers. (1975). United States Naval Technical Training Command: *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.* Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a006655.pdf

Conklin, Kathy and Norbert Schmitt. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics* 32: 45–61.

Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. (2000-2005). *Santa Barbara corpus of spoken American English, Parts 1-4*. Philadelphia: Linguistic Data Consortium.

Evert, Stefan. (2009). Corpora and collocations In A. Lüdeling and M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*, article 58. Berlin: Mouton de Gruyter.

Flesch, Rudolf. (2016, August 1). *How to Write Plain English.* Retrieved from http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml

Hallin, Daniel C. (1992). Sound Bite News: Television Coverage of Elections. In *Journal of Communication*, Spring 1992.

Levy, Roger and T. Florian Jaeger. (2007). Speakers optimize information density through syntactic reduction. In *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*.

Lim, Elvin T. (2008). *Anti-Intellectual Presidency: The Decline of Presidential Rhetoric from George Washington to George W. Bush*. Cary: Oxford University Press.

Lu, Xiaofei. (2014). *Computational methods for corpus annotation and analysis*. New York: Springer.

Malvern, David D., Brian J. Richards, Ngoni Chipere & Pilar Durán. (2004). *Lexical Diversity and Language Development*, Houndmills, UK: Palgrave MacMillan.

Pecina, Pavel. (2009). *Lexical Association Measures: Collocation Extraction*. (Studies in Computational and Theoretical Linguistics 4). Prague: Institute of Formal and Applied Linguistics, Charles University in Prague.

Peters, Gerhard and John T. Wooley. (2016, July 20). *Presidential Debates*. Retrieved from http://www.presidency.ucsb.edu/debates.php

Presidential Election Process. (2016, August 1). Retrieved from https://www.usa.gov/election#item-37162

Ryan, Kim. (2016, July 20). *Lingua::EN::Fathom*. Retrieved from http://search.cpan.org/~kimryan/Lingua-EN-Fathom-1.18/lib/Lingua/EN/Fathom.pm