JOURNAL OF QUANTITATIVE LINGUISTICS, 2017
https://doi.org/10.1080/09296174.2017.1349358

Routledge
Taylor & Francis Group

Trump's and Clinton's Style and Rhetoric during the 2016 Presidential Election

Jacques Savoy

Computer Science Department, University of Neuchâtel, Neuchâtel, Switzerland

ABSTRACT

The present paper examines the style and rhetoric of the two main candidates (Hillary Clinton and Donald Trump) during the 2016 presidential election. Based on interviews and TV debates, the most frequent lemmas indicate an emphasis on the pronoun *I* for both candidates while in speeches, the pronoun *we* appears more frequently. According to overall stylistic indicators, Trump adopts a simple and direct communication style, preferring short sentences, avoiding complex formulations and employing a reduced vocabulary. In the oral form, Trump frequently uses verb phrases (verbs and adverbs) and pronouns while Clinton is more descriptive (more nouns and prepositions). As expected, the speeches present differences from the oral form. For Trump, the difference is clearly larger, distinctively depicting two communication styles (oral and written). The specific terms or sentences associated with each candidate reveal their characteristic topics and style, such as the repetition of expressions and negativity for Trump. Based on predefined word lists, this study indicates that Clinton's rhetoric employs more cognitive words, while negative emotions and exclusive terms occur more frequently in Trump's verbiage.

1. Introduction

The 2016 US presidential election was characterized by two figures, both unloved by the majority of Americans (Yourish, 2016). Ignoring every norm of American politics and hoping to reflect the silent majority, Donald Trump says what he thinks, and thus appears sincere and authentic. For him, any exposure in and all comments from the media are considered good. His campaign has used social networks in a provocative way without any real consideration of the media (Boyd, 2016). Trump believes that the repetition of a simple message, even a wrong one (Millbank, 2016), is enough to persuade citizens that it is true. His image is centred around his verbosity, egocentricity, and pomposity. Just after the announcement of his candidacy for President (16 June 2015), his candidacy was mainly viewed as marginal, without any real future. But Trump

CONTACT Jacques Savoy Jacques.Savoy@unine.ch

was able to beat all his opponents, won the nomination for the Republican party (21 July 2016), and won the general election (8 November 2016).

Nominated by the Democrats (28 July 2016), Hillary Clinton always appeared as a cold woman, somewhat robotic, and moreover as a member of a political establishment rejected by many people. She doesn't like the press and, in return, it doesn't like her much either. This aspect might be related to her earlier years at the White House as an overqualified First Lady who wanted to play a principal role in politics (e.g. the Clinton health care plan in 1993). For some people, she is even a crook and a liar, or, at least, dishonest (Sainato, 2016). When her campaign started (14 April 2015), everything seemed simple and the road to the nomination appeared to be set without any real problems. The presence of Sanders occupying a position more on the left demonstrated that the Democratic primaries were more difficult than expected. Finally, her email case and FBI investigations were a real concern for her image in the public eye, especially during the general election campaign.

Based on TV debate transcripts, interviews, and speeches delivered during the electoral campaign, can we detect their communication style and rhetoric? Can we discover the rhetoric features that can explain Trump's or Clinton's success or, at least, their differences? Can we measure the stylistic distance between the candidates in both parties? To provide a partial answer to these questions, we define rhetoric as the art of effective and persuasive speaking, the way to motivate an audience, while language style is presented as pervasive and frequent forms used by an author or speaker (Biber & Conrad, 2009).

The rest of this paper is organized as follows. The next section exposes some related research in computer-based analysis of political speeches. The third section briefly presents some statistics about our corpus. The fourth describes and applies different measurements and methods to define the rhetoric and style of the different candidates. Based more on topics, the fifth visualizes the relative position of each candidate in different spaces. A conclusion presents the main findings of this study.

2. Related Work

Political texts have been the subject of various studies discussing different aspects. Focusing on governmental speeches, Labbé and Monière (2003, 2008a) have created a set of governmental corpora such as the *Speeches from the Throne* (Canada and Quebec), a corpus of the general policy statement of French governments, as well as a collection of press releases covering the 2012 French presidential campaign (Arnold & Labbé, 2015; Labbé & Monière, 2013). Similar research has been conducted with other languages, such as Italian (Pauli & Tuzzi, 2009). From these analyses, we can observe, for example, that governmental institutions tend to smooth out the differences between political parties when exercising power. Moreover, the temporal period constitutes an important

factor in explaining the variations between presidents or prime ministers. The presence of a strong leader is usually accompanied by a real change in the style and the vocabulary of governmental speeches (Labbé & Monière, 2003).

Focusing on the United States, recent studies confirm these findings, as for example, based on the *State of the Union* (Savoy, 2015) or inaugural addresses (Kubát & Cech, 2016). Time plays a more important factor than political affinities in stylistic variations, as well as exceptional events (e.g. worldwide war, deep economic depression).

Differentiations between political parties can however be observed as, for example, based on tweets (Sylwester & Purver, 2015). Such differences tend to be correlated with psychological factors. For example, positive emotion words occur more frequently in Democrats' tweets than in Republican ones, as well as swear expressions, or first singular person pronouns (e.g. *I*, *me*). In a related study based on a training corpus, Laver, Benoit, and Garry (2003) describe a methodology for extracting political positions from texts. In a similar vein, Yu (2008) demonstrates that machine learning methods (e.g. SVM and naïve Bayes) can be trained to classify congressional speeches according to political parties. Better performance levels can be achieved when the training examples are extracted from the same time period as the test set. In another study, Yu (2013) reveals that feminine political figures tend to use emotional words and personal pronouns more frequently than men. A more general overview of using different computer-based strategies to detect and extract topical information from political texts can be found in Grimmer and Stewart (2013).

Web-based communication (e.g. tweets, blogs, chats) was used by O'Connor, Balasubramanyan, Routledge, and Smith (2010) to estimate the popularity of the Obama administration. This study found a positive correlation between the presidential approval polls and positive tweets containing the hashtag #obama. Such a selection strategy produces a low recall (many tweets about Obama's administration are not considered). As a tweet is rather short (in mean eleven words), the sentiment estimation is simply the count of the number of positive and negative words appearing in the OpinionFinder dictionary (Wilson et al., 2005).

Based also on several word lists, Young and Soroka (2012) describe how one can detect and measure sentiments appearing in political texts. The suggested approach is rather similar to that of O'Connor et al. (2010), counting the frequency of occurrence of words appearing in a dictionary of positive or negative emotion terms. Using different lists of words, Hart (1984) has designed and implemented a political text analyser called DICTION. Based on US presidential speeches, Hart (1984) presents the rhetoric and stylistic differences between the US presidents from Truman to Reagan, while a more recent study (Hart, Childers, & Lind, 2013) exposes the stylistic variations from G.W. Bush to Obama. Using the DICTION system, Bligh, Merolla, Schroedel, and Gonzalez (2010) analyse the rhetoric of H. Clinton during the 2008 presidential election.

Hillary appears more feminine than the other candidates, using more *I* than *we*, and showing a higher frequency in the category *Human Interest* (e.g. family, man, person, etc.).

As another example, LIWC (Linguistic Inquiry and Word Count) (Tausczik & Pennebaker, 2010) regroups different categories used to evaluate the author's psychological status (e.g. feminine, emotion, leadership), as well as her/his style (e.g. based on personal pronouns (Pennebaker, 2011). The underlying hypothesis is to assume that the words serve as guides to the way the author thinks, acts or feels. In LIWC, the generation of the word lists was done based on the judgments of three experts instead of simply concatenating various existing lists. Using the LIWC system, Slatcher, Chung, Pennebaker, and Stone (2007) were able to determine the personalities of different political candidates (2004 US presidential election). They defined the psychological portrait both on single measurements (e.g. the relative frequency of pronouns, social words, etc.) and using a set of composite indices reflecting the cognitive complexity, presidentiality or honesty of each candidate. These personality measurements were in agreement with different opinion polls. For example, G.W. Bush uses the pronoun *I*, positive emotion words (e.g. happy, truly, win), and future tense more frequently. The public perceives J. Kerry as a kind of depressed person, serious, sombre, and cold, uttering negative emotion expressions (e.g. sad, worthless, cut, lost) and physical words (e.g. head, ache, sleep) more frequently.

In brief, previous studies have mainly analysed governmental speeches, and less frequently the electoral speeches (Boller, 2004) or related messages (such as press releases (Labbé & Monière, 2013). A few studies focus on the legislative level (e.g. the Congress) and these studies are mainly based on the written form. More recently, the web-based communication channels have been studied, but in this perspective based on tweets and less frequently on blogs, or audio and video media (e.g. YouTube). The present study focuses on two less explored aspects, namely the electoral campaign on the one hand, and on the other, the oral form.

3. Electoral Corpus

To analyse the rhetoric and style adopted by the two nominees during the 2016 US presidential election, the transcripts of the TV debates during the primaries were downloaded from the Internet (mainly from the website www.presidency.ucsb.edu). For the Republican candidates, twelve TV debates were organized, from the first one held on 6 August 2015 with 10 candidates to the last one organized 10 March 2016 with four candidates. For the Democrats, one can count nine TV debates held from 13 October 2015 (with five candidates) to 9 March 2016 with two candidates. In addition to these transcripts, we have also considered 27 interviews given by Clinton and two by Trump. This first set of documents forms our first oral corpus denoted 'Clinton Oral' or 'Trump Oral'.

As a second corpus, the transcripts of the three presidential TV debates have been included to complement this study (26 September, 9 October, 19 October 2016). The audio source was not used directly, but rather the transcripts of the debates between the two candidates. From this textual representation, the processing can be done as for usual written speeches or messages. This corpus denoted 'Pres. debates' also corresponds to an oral text genre. This second set serves mainly to establish some contrasts with our first oral corpus.

The third corpus represents a written communication genre. It is composed of 37 speeches uttered by Clinton and 58 speeches given by Trump. The label 'Speeches' is used to designate this corpus. One can argue that electoral speeches delivered by the candidates match more an oral communication genre. However, as mentioned by Biber and Conrad (2009, p. 262):

> Language that has its source in writing but performed in speech does not necessarily follow the generalization [written versus oral]. That is, a person reading a written text aloud will produce speech that has the linguistic characteristics of the written text. Similarly, written texts can be memorized and then spoken.

To illustrate the differences between the two forms of communication, the following examples indicate clearly that the TV debates represent the oral form while the passage extracted from an electoral speech corresponds more to a written form even if finally it is uttered.

> I've been challenged by so many people, and I don't frankly have time for total political correctness. And to be honest with you, this country doesn't have time either. This country is in big trouble. We don't win anymore. We lose to China. We lose to Mexico both in trade and at the border. We lose to everybody. (D. Trump, Republican candidates debate, Cleveland, OH, 6 August 2015)

> We are also going to have to change our trade, immigration and economic policies to make our economy strong again and to put Americans first again. This will ensure that our own workers, right here in America, get the jobs and higher pay that will grow our tax revenue and increase our economic might as a nation. (D. Trump, speech, National Press Club, Washington, DC, 27 April 2016)

Besides the text genre difference, this collection is relatively homogenous, corresponding to text extracted from a short period of time, with the same main objectives (convincing the people, answering questions, presenting candidate's ideas and solutions). Several factors influencing the style are therefore fixed. Thus, the remaining variations can be largely explained by the text genre, the author, and topical variations.

Table 1 reports the vocabulary size (number of distinct word types) for each corpus and candidate as well as the total number of word tokens (text size). To understand the difference between a word type and a token, consider the following sentence: 'the law is harsh, but it is the law'. One can count nine word tokens (or simply tokens) and six word types (or types). Ignoring the punctuation, the type 'the', 'is', or 'law' occurs twice. The set of all distinct word types forms the vocabulary, denoted by V, while the text size is represented

Table 1. Some statistics about the three corpora according to the two candidates.

	Trump, Oral	Clinton, Oral	Trump, Pres. debates	Clinton, Pres. debates	Trump, Speeches	Clinton, Speeches
Vocabulary	3,027	5,329	1,953	2,077	6,761	6,865
Tokens	63,589	124,749	25,059	15,961	166,111	140,538

by *n*. As shown in Table 1, the first and last corpus represent the largest ones. These two corpora form the main ground for our investigations and findings.

4. Evaluation of Stylistic Characteristics of the Candidates

To discriminate between the different styles adopted by the candidates, Biber and Conrad (2009) indicate that such a study should be based on ubiquitous and frequent forms. Thus, the analysis of the most frequent terms is a good starting point, as shown in the first of the following subsections. The second subsection proposes considering four overall stylistic measurements and applies them to the different candidates. The last subsection describes the differences in the distribution of the grammatical categories between candidates.

4.1. Most Frequent Lemmas

Our first quantitative linguistics study focuses on word occurrence frequencies. As the English language has a relatively simple morphology, considering the inflected forms (e.g. *we*, *us, ours*, or *wars*, *war*) or the lemmas (dictionary entries such as *we* or *war*) leads to similar conclusions. This latter form is mainly exploited in the current study.

To define the corresponding lemma (and POS) of each token, the Part-Of-Speech (POS) tagger proposed by Toutanova, Klein, Manning, and Singer (2003) was applied. For each sentence given as input, this system provides the corresponding POS tag for each token. For example, from the sentence 'Our energy policy is creating new jobs.' the POS tagger returns 'Our/PRP$ energy/NN policy/NN is/VBZ creating/VBG new/JJ jobs/NNS ./.'. Tags may be attached to nouns (NN – noun, singular; NNS – noun, plural), verbs (VB – base form; VBG – gerund or present participle; VBZ – third-person singular present), adjectives (JJ), personal pronouns (PRP), prepositions (IN), determiners (DT) and adverbs (RB). With this information, we are then able to derive the lemma by removing the plural form of nouns (e.g. *jobs*/NNS → *job*/NN) or by substituting inflectional suffixes of verbs (e.g. *creating*/VBG → *create*/VB). Finally, this POS tagger defines the sentence boundaries used in our analysis.

Our first analysis considers the most frequent lemmas occurring in the oral and written speeches. Unsurprisingly, the article *the* and the verb *be* (lemmas of the type *am, is, are, was*, etc.) appear regularly in the first two ranks. Looking at

Table 2. The top ten most frequent lemmas according to our three corpora.

Trump, Oral	Clinton, Oral	Trump, Pres. debates	Clinton, Pres. debates	Trump, Speeches	Clinton, Speeches
be	*be*	*be*	*be*	*the*	*be*
I	***I***	*the*	*the*	*be*	*to*
the	*the*	***I***	*to*	*and*	*and*
to	*to*	*and*	***we***	***we***	*the*
and	*and*	***we***	*and*	*to*	***we***
we	*that*	*to*	***I***	*of*	***I***
have	***we***	*have*	*that*	*an*	*of*
an	*have*	*you*	*have*	***I***	*an*
it	*of*	***it***	*of*	*in*	*that*
that	*an*	*of*	*an*	*have*	***you***

the most frequent lemmas in the Brown corpus (Francis & Kucera, 1982), the first two are the same, but after that the order changes. In the Brown corpus, the top ten most frequent lemmas are *the*, *be*, *of*, *and*, *to*, *a*, *in*, *he, have*, and *it.*

Table 2 reports the top ten most frequent lemmas for each of the two candidates and the three corpora. In this table, the personal pronouns are depicted in bold. As one can see, the first-person pronoun (***I*** or ***we***) appears relatively high in this list (and does not appear in the top ten positions in the Brown corpus) and, to a lesser extent, *you* (Clinton's speeches) or *it* (Trump Oral). Even if pronouns are more frequent in dialogue or in oral form than in written communication (Biber, Conrad, & Leech, 2002), the high frequency of *I* and *we* is a fundamental characteristic of the political speech. In the Brown corpus, only *he* and *it* appear in the top ten most frequent lemmas. Moreover, usually a high frequency of *we* is associated more with speeches of a president or a prime minister (Pennebaker, 2011). For example, in Obama's *State of the Union* addresses, the frequencies of the article *the* and the pronoun *we* are very similar. In a related study, Bligh et al. (2010) found that Clinton changed her voice during the 2008 presidential election. One of the most important adjustments was a higher use of *I* and *me* and a decrease in the occurrence of the pronoun *we*. For Labbé and Monière (2008b) too, the pronoun *I* was clearly over-used during the 2007 French presidential election.

This table reveals another interesting fact related to the frequencies of the pronouns ***we*** and ***I***. The written form tends to use ***we*** more frequently than ***I***. The pronoun ***we*** owns the useful advantage of being ambiguous. (Who is really behind the ***we***? 'Myself and the future government?' 'Me and the people?' 'Me and the workers?' 'Me and the [future] Congress?', etc.)

4.2. Global Stylistic Measurements

In various stylometric studies, different overall measurements have been proposed to quantify the lexical and syntactic choice of the author, as well as to provide a complexity value for the underlying text. As a first indicator, the

Table 3. Four global stylistic measurements over the three corpora and the two candidates.

	Trump, Oral	Clinton, Oral	Trump, Pres. debates	Clinton, Pres. debates	Trump, Speeches	Clinton, Speeches
MSL	*13.3*	19.6	14.1	18.6	17.8	**20.1**
LD (%)	*36.8*	39.3	37.5	40.5	**46.4**	41.8
BW (%)	*18.3*	22.9	20.3	23.8	**29.1**	23.8
TTR	*29.3*	35.3	30.3	35.8	**39.6**	37.3

mean sentence length (number of tokens per sentence) reflects a syntactical preference. Longer sentences are more complex to understand, especially in the oral communication form. Based on the *State of the Union* addresses given by the Founding Fathers, this average value is 39.6 (with Madison depicting the highest mean sentence length with 44.8 tokens/sentence). With Obama, the mean sentence length decreases to 18.5 tokens/sentence. These examples clearly indicate that the style is changing over time. Nowadays, the stylistic norm prefers shorter formulations, easier to understand for the whole audience.

As shown in Table 3 under the row labelled 'MSL' (Mean Sentence Length), Trump prefers uttering short sentences (Oral: 13.3 tokes/sentence) while Clinton adopts longer formulations without a real difference between the oral form (19.6) and the written (20.1). Both values are higher than Obama's average (18.5). The presence of long sentences indicates a substantiated reasoning or specifies the presence of a more detailed explanation. Even if a long sentence is required, its length does not guarantee clear understanding. Of course, with the written genre, the mean sentence length tends to increase (e.g. Trump from 13.3 to 17.8). Looking back to our two examples in Section 3, one can observe that, even if two passages have a similar length, the oral one is composed of seven sentences while the written one contains only two. In Table 3, the largest value per row is depicted in bold, and the smallest value in italics and always appears under the column 'Trump, Oral'.

The Lexical Density (LD) corresponds to a second measurement employed to indicate the informativeness of a text. The underlying computation is depicted in Equation (1), where the variable $n(t)$ indicates, for a text t, the total number of tokens (or text length), *function words*(t) the number of function words in t, and *lexical words*(t) the number of lexical words in t. This latter set is composed of nouns, names, adjectives, verbs, and adverbs. On the other hand, function words regroup, by definition, all other grammatical categories, namely determiners (e.g. *the*, *this*), pronouns (*I*, *us*, …), prepositions (*to*, *in*, …), conjunctions (*and*, *but*, …), modal verbs, and auxiliary verb forms (*has*, *would*, *can*, …). The list of functional words for the English language contains 402 forms.

$$LD(t) = \frac{\textit{lexical words}(t)}{n(t)} = 1 - \frac{\textit{function words}(t)}{n(t)}. \tag{1}$$

A relatively high *LD* percentage indicates a more complex text, containing more information. When comparing the oral and written genre, the latter tends to present a higher lexical density. This relationship can be found in Table 3 for both candidates (Trump: 36.8 versus 46.4%; Clinton: 39.3 versus 41.8%). For this measurement and for the oral genre, one can observe that Trump uses more functional words while Clinton provides more information (or more lexical terms). As for the MSL, the difference between the oral and written form is relatively small when analysing Clinton's style but this is not the case with Trump.

As an additional global stylistic measurement, the frequency of big words (composed of six letters or more, and denoted BW) is shown in Table 3. A text or a dialogue with a high percentage of big words is more complex to understand, as indicated by Lakoff and Wehling (2012):

> One finding of cognitive science is that words have the most powerful effect on our minds when they are simple. The technical term is basic level. Basic-level words tend to be short. […] Basic-level words are easily remembered; those messages will be best recalled that use basic-level language.

This rhetoric problem was recognized by previous US presidents such as President Johnson, who stated to his ghostwriters: 'I want four-letter words, and I want four sentences to the paragraph' (Hart, 1984). The smallest percentage of big words can be found with Trump (18.3%) compared to 22.9% for Clinton (oral genre). With the written speeches, this mean value increases to 23.8% for Clinton, and 29.1% for Trump, the highest value for this indicator.

The TTR (Type–Token Ratio) or the relationship between the vocabulary size and the number of word types (Baayen, 2008; Mitchell, 2015) corresponds to our last global stylistic measure. A high value indicates the presence of a rich vocabulary showing that the underlying text exposes many different topics or that the author presents a theme from several angles with different formulations. To compute this value, one divides the vocabulary size (number of types) by the text length (number of tokens). This estimator has the drawback of being unstable, tending to decrease with text length (Baayen, 2008). To avoid this problem, the computation applied in this study is provided in Covington and McFall (2010) or Popescu (2009), who suggest taking the moving average.

From the data depicted in Table 3, one can see that the TTR value reaches a minimum of 29.3 (Trump, Oral) to a maximum of 39.6 (Trump, Speeches). As for the other measurements, a higher value can be expected when analysing a written text compared to an oral one. This ratio is respected by values depicted in Table 3. When comparing Trump's and Clinton's choices, one can see that, in the oral form, Trump's style is simpler, reusing the same words and expressions more often than Clinton's.

The general trends that can be extracted for Table 3 are the following. One can see two Trump figures, one related to the oral mode (columns 'Oral' and 'Pres. debates'), the second in written speeches. Trump's oral genre is simple, direct, using short sentences (MSL) and fewer hard-to-understand words

(BWs), preferring to repeat the same terms (TTR). In all measurements computed for the oral genre, Trump's values are smaller than Clinton's. In the written form, Trump opts for richer expressions and formulations (*LD*: oral 36.8 versus written 46.4%), more complex lexical choices (BW: 18.3 versus 29.1%), and a reduction in repetition (higher TTR values: 29.3 versus 39.6). For Trump, the high number of differences between the oral and written genre clearly indicates the presence of a ghostwriter (or a team thereof) without important modifications done by the Republican nominee. When comparing Clinton's oral and written genre, differences do exist but they are small, reflecting that the candidate is writing her own speeches or, at least, has a close control over her writers. Compared to Trump, Clinton's oral form is usually more complex, based, on average, on longer sentences (MSL) and having a higher informativeness value (*LD*). Her lexical choice reflects a richer vocabulary (TTR) with more complicated words (BWs).

4.3. Part-of-Speech Distribution

To study the difference in style and rhetoric between the two candidates, the relative frequencies of the Part-Of-Speech (POS) or grammatical categories can provide useful information. Two main syntactic constructions can be chosen by a speaker, namely using verb phrases more frequently (composed of verbs and adverbs) or choosing noun phrases more often (with more nouns, adjectives, determiners, and prepositions). To analyse this aspect, Table 4 presents the POS distributions, as percentages, over the three corpora and two candidates. The maximum value per grammatical category is shown in bold, and the minimum in italics.

The data depicted in Table 4 indicate that, in the oral form, both candidates employ verb constructions more frequently (Trump verb, oral: 25.8 versus 21.0% in written speeches) while nouns, adjectives, and conjunctions occur more often in the written speeches (Clinton noun, oral: 15.4 versus 17.4% in messages). Moreover, pronouns are used more intensively in dialogue and in

Table 4. POS distribution according to our three corpora and two candidates.

	Trump, Oral (%)	Clinton, Oral (%)	Trump, Pres. debates (%)	Clinton, Pres. debates (%)	Trump, Speeches (%)	Clinton, Speeches (%)
Noun	*13.4*	15.4	14.4	16.5	**19.3**	17.4
Name	4.2	4.2	4.5	4.1	**7.1**	*4.0*
Pronoun	**16.1**	13.2	15.2	13.1	*10.2*	12.7
Adjective	*5.4*	5.9	5.7	6.0	**7.4**	6.9
Verb	**25.8**	24.3	25.0	24.0	*21.0*	22.1
Adverb	**8.8**	7.3	7.9	6.9	*5.5*	7.2
Determiner	9.0	9.5	8.9	9.5	**9.7**	*8.7*
Preposition	*11.8*	**15.1**	12.4	14.4	13.6	14.7
Conjunction	*3.7*	4.0	4.2	3.8	4.1	**4.9**
Other	1.7	1.1	1.8	1.6	2.1	1.4
Q-index	**82.8**	81.1	81.3	80.0	*73.9*	76.4

oral communication. Therefore, the data depicted Table 4 confirm some of the differences found between the oral and written genres (Biber et al., 2002) as, for example, the more frequent use of verbs and pronouns in dialogue. In our case, Trump presents a pronoun frequency of 16.1% in the oral form versus 10.2% in written speeches.

When contrasting the two candidates based on the oral genre, Clinton's interventions use noun constructions more frequently (nouns, 15.4%; adjectives, 5.9%; determiners, 9.5%) and clearly more prepositions (15.1%). In Trump's oral form, one can observe more verbs (25.8%) and adverbs (8.8%) as well as more pronouns (16.1%). Comparing both nominees in the TV presidential debates, a similar finding is apparent. When considering the written genre, the POS distribution for Trump is clearly different from those represented during the primaries. Compared to Clinton, Trump uses nouns and determiners more often, favouring noun phrases in his speeches. Moreover, Trump's writings contain more names (7.1%) (e.g. Mexico, China, Clinton, …), anchoring his remarks more in space and in relation to people. Compared to Trump, Clinton employs verbs slightly more often, and clearly more adverbs and pronouns.

To obtain an overall measure of the intensity of the action over the descriptive part of a text, Kubát and Cech (2016) suggest computing the ratio between the proportion of verbs divided by the sum of the proportion of the verbs and adjectives as depicted in Equation (2).

$$Q\text{-index} = \frac{\text{percentage of verbs}}{(\text{percentage of verbs} + \text{adjectives})}. \tag{2}$$

The underlying idea is to quantify the activity by verbs while the descriptiveness of a text is represented by the proportion of adjectives. The values of this Q-index are depicted in the last row of Table 4. Distinctively, the form depicted during the primaries presents a higher Q-index value compared to the written genre, which is more descriptive. In the primaries, Trump shows a higher value than Clinton, indicating a communication marked with more action. The difference is however not very large.

5. Evaluation of Topical Characteristics of the Candidates

The previous section focused mainly on stylistic features, both at the lexical and syntactical level. When looking more at the themes, one can also observe differences between the candidates. In the first of the following subsections, a textual distance is presented and used to derive graphs representing either the stylistic or topical affinities between the candidates. Then, a technique for defining the specific terms and sentences for each nominee is described and some examples are given. The last subsection analyses the two candidates based on a set of dictionaries to reveal some of their rhetorical differences.

5.1. Stylistic and Topical Distance Between the Candidates

The oral and written forms of communication present stylistic differences as shown previously. Instead of comparing such variations one by one, we propose to compute a distance reflecting their similarities and divergences more globally. To achieve this, a text is viewed as a composite object containing the style with its lexical, syntactical, or discourse factors, and words belonging to the thematic aspects. To split according to these two main components, the first map is generated according to the stylistic aspects while the second takes into account only the topical elements.

To reflect the style, various studies have based their findings on functional words (e.g. determiners, pronouns, prepositions, pronouns, conjunctions, and auxiliary verb forms). As these grammatical categories are closed (one cannot generate a new word in this set), we enumerate all possible forms for the English language and create a list with 402 entries. Thus, when considering the stylistic aspect of a given text, only these words are counted.

To define a distance between Text A and Text B, Equation (3) (Labbé, 2007) is applied, in which V_A (or V_B) indicates the vocabulary of Text A, tf_{iA} (respectively tf_{iB}) denotes the term occurrence frequency of the ith word type in Text A, and n_A (respectively n_B) the length of Text A (number of tokens).

$$\text{dist}(A, B) = \frac{\sum_{i \in V_A \cup V_B} |tf_{iA} - tf_{iB}|}{2 \cdot n_A}. \tag{3}$$

This formulation assumes that both texts have the same length ($n_A = n_B$). This is however rarely the case, and one needs to reduce the largest text (assuming it is Text B) to the size of the smallest one (Text A in our example). To achieve this, the term frequency of each word type belonging to the largest text is modified as follows:

$$tf'_{iB} = tf_{iB} \cdot \frac{n_A}{n_B}. \tag{4}$$

Based on this measure, one can display directly the 6×6 matrix containing these distances for the two nominees and the three corpora. To obtain a better visualization, a clustering method – e.g. hierarchical clustering based on the complete link (Baayen, 2008) – can be applied to regroup the candidates sharing similarities. Such distance matrices can also be represented by a tree-based visualization method (Baayen, 2008; Paradis, 2011). Following this approach, Figure 1 illustrates the differences based only on the style. In such graphs, not all distances are fully respected. The visualization algorithm produces the best possible two-dimensional representation, trying to respect as best as possible the real distances between all points. Some deformations are however always present.

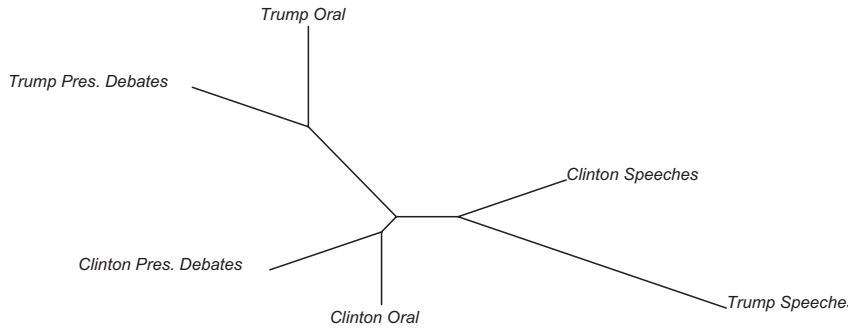

Figure 1. Stylistic distance between the two candidates and the three text genres.

In this figure, the distance between two end points is indicated by the length of the lines needed to connect them. For example, starting with 'Clinton, Speeches', one can follow the branch until reaching the central point, then we can go along the lines leading to the targeted point (e.g. 'Clinton, Oral').

In Figure 1, the closest distance (0.083) is between Clinton's presidential debates and Clinton's oral transcripts, while the second smallest (0.10) links Trump's presidential debates and Trump's oral form. Considering the TV presidential debates on the one hand and, on the other, the TV primary debates and the interviews, this figure indicates that the oral style in these two contexts is similar for both candidates. The third smallest distance can be found between Clinton's speeches and Clinton's presidential debates. As shown previously, Clinton's communication style forms a relatively homogenous entity. On the contrary, the longest distance (0.245) connects Trump's speeches and Trump's oral form, and the second longest (0.209) links Trump's speeches to Trump's presidential debates. Obviously, Trump has adopted two distinct styles for the oral and written communication channels. This finding confirms the presence of a (or a team of) ghostwriter(s) less supervized by the Republican nominee.

To generate Figure 2, the intertextual distance is computed according to topical words. To achieve this, the computation ignored all functional words for all texts. In this figure, the closest link (0.289) can be found between Trump's presidential debates and Trump's oral form, while the second closest (0.293) links Clinton's oral form and Clinton's speeches. The third smallest distance (0.333) is between Clinton's oral form and Clinton's presidential debates. The longest distance (0.435) connects Trump's presidential debates to speeches given by Clinton, and the second longest connects (0.431) the speeches uttered by Trump to Trump's oral form.

In Figure 2, one can see that Clinton's remarks in the oral and written genres are relatively close together. For Trump, the two oral points are relatively near, but a higher distance links the oral with the written form. The topical words in the two text genres are therefore less similar than with Clinton.

5.2. Most Specific Terms

During an electoral campaign, each candidate wants to promote his/her own specific point of view on the most important issues, and tries to underline his/her differences with the other's. Just considering the top most frequent words, similar sets appear with each candidate. For example, Trump and Clinton prefer using the pronoun *I* instead of *we* in oral communication, while in their speeches they use *we* more frequently (see Table 2).

To measure the specificity attached to a term (Muller, 1992), the corpus is divided into two distinct parts denoted P_0 and P_1. For a given term t_i, its occurrence frequency in P_0 is given by tf_{i0}, and in P_1 by tf_{i1}. In this study, P_0 corresponds to all comments uttered by a given candidate, while P_1 denotes all other comments and remarks. Thus, for the entire corpus the occurrence frequency of the term t_i is $tf_{i0} + tf_{i1}$. The total number of lemmas in part P_0 (or its length) is denoted n_0, similarly with P_1 and n_1, and the length of the entire corpus is defined by $n = n_0 + n_1$.

For any term t_i, we assume that its distribution follows a binomial, with parameters n_0 and $p(t_i)$ representing the probability of the term t_i being randomly selected from the entire corpus. Based on the maximum likelihood principle, this probability is estimated as $p(t_i) = (tf_{i0}+tf_{i1})/n$.

Through repeating this drawing n_0 times, the expected number of occurrences of term t_i in P_0 can be estimated by $n_0 \cdot p(t_i)$. This value is then compared with the observed number (namely tf_{i0}) and a large difference between these two values indicates a deviation from the expected behaviour. To obtain a more precise definition of *large* we account for the binomial variance (defined as $n_0 \cdot p(t_i) \cdot (1-p(t_i))$). Equation (5) defines the final standardized Z-score (or standard normal distribution $N(0,1)$) for term t_i, using the partitions P_0 and P_1.

$$Z\text{-score}(t_{i0}) = \frac{tf_{i0} - n_0 \cdot p(t_i)}{\sqrt{n_0 \cdot p(t_i) \cdot (1 - p(t_i))}}. \tag{5}$$

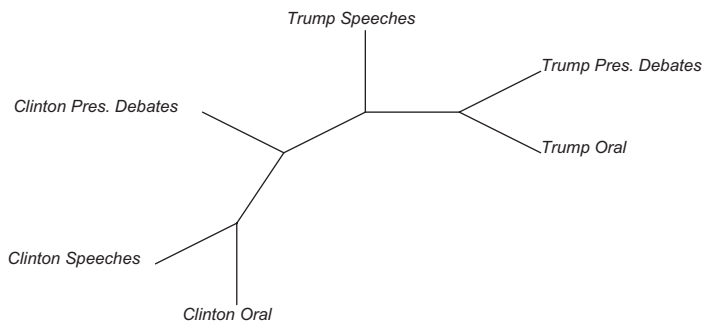

Figure 2. Topical distance between the candidates.

Table 5. The top ten most specific terms per candidate.

Trump, Oral	Clinton, Oral	Trump, Pres. debates	Clinton, Pres. debates	Trump, Speeches	Clinton, Speeches
I	I	she	Donald	Clinton	families
not	think	Mosul	he	Hillary	work
do	Senator	bad	determines	will	and
very	that	outsmarted	pregnancy	American	kid
Jeb	Sanders	leaving	cyberattack	she	here
ok	know	you	minutes	job	young
it	well	do	quoting	America	disabilities
he	Republican	Russia	avoid	illegal	to
excuse	what	Lester	undocumented	nation	you
Ted	comprehensive	look	discipline	foreign	he

Applying this procedure, the term specificity can be computed according to the text P_0. These Z-score values can verify whether the underlying lemma is used proportionally with roughly the same frequency in both parts (Z-score value close to zero). A positive Z-score larger than a fixed threshold (e.g. 3) indicates that the term is *significantly over used*, and belongs to the specific vocabulary of P_0. In other words, the text P_0 contains significantly more occurrences of the corresponding term than expected by a uniform distribution over the whole corpus. A large negative Z-score (less than $-\delta$) indicates that the corresponding term is significantly under used in P_0.

Table 5 shows the top ten most over-used terms for both candidates. One can see the presence of expressions related to the dialogue between candidates (*Jeb* [Bush], *Ted* [Cruz], *Lester, Hillary, Clinton* as well as *she* with Trump; *Senator*, *Sanders*, *Donald*, *he* with Clinton).

A more interesting finding is the presence of the pronoun *I* in the most over-used terms in the oral corpus for both Trump and Clinton. A candidate who wants to stay in the running must put himself/herself forward. After all, the election is a procedure to select a leader. During the interviews and TV debates of the primaries, both Clinton and Trump have clearly put forward their personas. Of course, behind a candidate a political programme must also appear. Some of the terms depicted in Table 5 give some indications of this aspect, for example, *Mosul, Russia, job, illegal* with Trump, or *pregnancy, cyberattack, families, work, kid, young, disabilities* with Clinton.

5.3. Most Specific Sentences

Listing the most over-used terms is sometimes not enough to have a clear understanding of the candidate's position on a given issue. To obtain a more precise idea, one possible approach is to extract a reduced set of specific sentences from each candidate. Such a sentence is defined as one having the largest number of specific terms.

Based on this definition, our system has extracted the following specific sentences per candidate in which the most over-used terms are depicted in italics. Starting with the TV presidential debates, the following are the representative sentences delivered by Trump.

> It turned out *she did say* the *gold* standard and *she said she didn't say* it. (D. Trump, second presidential debate, 9 October 2016)

> So what *they are* doing *is leaving* our country and, believe *it* or *not*, *they are leaving because taxes are* too high and *because* some of them *have* lots of *money* outside of our country and instead of bringing *it* back and putting the *money* to work *because* they can*'t* work out a deal and everybody agrees *it* should *be* brought back, instead of that, *they are leaving* our country to get their *money because they* can*'t* bring their *money* back into our country *because* of bureaucratic red *tape*, *because they* c*an't* get together. (D. Trump, first presidential debate, 26 September 2016)

As usual in a debate and illustrated in our first example, each candidate must demonstrate his/her credibility and tries to show that his/her opponent is a liar or an inconsistent person. The second sentence also illustrates some of Trump's stylistic aspects. Clearly Trump reuses the same words again and again (*because* occurs six times, *money* four times, as well as the negation *not*). The repetitions appear also at the syntactical level (many clauses starting with *because they* [*can't, are*]).

In a similar vein, the first sentence from Clinton's presidential TV debates is an attack against Trump. The second one indicates that the topics related to jobs are important (and even essential) during an election (and the fact that the speaker is interrupted by his opponent).

> So I actually think the most important question of this evening, *Chris*, is, finally, will *Donald* Trump admit and condemn *that* the *Russians* are doing this and make it *clear that he* will not have the help of Putin in this election, *that he* rejects *Russian espionage* against Americans, which *he* actually *encouraged* in the past? (H. Clinton, third presidential debate, 19 October 2016)

> We have a very robust set of plans *and* people have looked at both of our plans, have *concluded that mine* would create 10 million jobs *and* yours would lose us three and a half million jobs *and* (H. Clinton, third presidential debate, 19 October 2016)

In the written speeches, the candidates employ more noun phrases and produce longer sentences. Starting with Trump, the next sentence indicates the possible consequences of the programme proposed by his opponent or the consequences of her previous actions:

> Here is a summary *of the Hillary plan*: support *for Sanctuary Cities*; Social Security, Medicare, and lifetime welfare *for illegal immigrants by* making them all *citizens*; *Obamacare for illegal immigrants*; no deportation *of visa* overstays; expanding catch-and-release on *the border*; expanding President Obama's unconstitutional executive amnesty, including instant work permits *for millions of illegal* workers;

> freeing even more *criminal aliens by* expanding Obama's non-*enforcement* directives; a 550% *increase* in Syrian *refugees*. Either *we* win *this* election, or *we* lose *the country*. (D. Trump, speech, 17 October 2016)

> *She* supported Bill *Clinton*'s *NAFTA*, *she* supported *China*'s entrance into the *World* Trade Organization, *she* supported the *job*-killing *trade* deal with South Korea, and *she* supports the *Trans-Pacific* Partnership. (D. Trump, Detroit Economic Club, 8 August 2016)

In this last example, one can see that Trump's written sentences can be short when explaining one of his objectives (but always with repetitions, in this case the word *America*).

> *American* cars will travel the roads, *American* planes will soar in the skies, and *American* ships will patrol the seas. (D. Trump, speech, 13 September 2016)

Clinton has clearly selected a more complex communication style based on longer sentences compared to Trump, with a richer vocabulary. Our system extracts the following example:

> Imagine if *you* believe the minimum *wage should* be a living *wage*, if *you* believe – if *you* believe that *we* finally *should* have *paid family* leave in this country like every other *advanced economy*, if *you* believe *climate* change is real and *we* could save *our* planet by creating a lot of jobs at the same time, if *you* believe diversity is America's strength, not a weakness, if *you* believe women *should* be able *to* make *our* own health care decisions and that LGBT Americans *should* be treated equally *across* America, and *you should* be able *to* live up *to your* potential, no matter who *you* are *or* where *you* come from, then start *voting* October the 12th! (H. Clinton, speech, 10 October 2016)

The candidate is repeating *if* six times to insist on a set of issues (the requirement of increasing the minimum wage, and her concerns about climate change, jobs, equality, and health care).

> *And* I *want* to also commend him *for* talking about his *mother and how hard* she *worked*, *and* his *sister* who *he* helped to raise. (H. Clinton, speech, 30 September 2016)

This last example clearly corresponds to a feminine author – who usually uses more pronouns and words related to family and relatives (Pennebaker, 2011). In this sentence, Clinton wants to present herself as a person belonging to the same group as the audience, a technique applied to increase the charisma of the leader (Bligh et al., 2010).

5.4. Semantic-Based Analysis

As another technique to detect and analyse variations in rhetoric, one can take advantage of some dictionaries proposed by Hart (1984) (the Diction system) or those suggested by Tausczik and Pennebaker (2010) (LIWC). These two text analysis systems regroup under one category words belonging to the same semantic or syntactic group. For example, under *Symbolism*, one can

Table 6. Percentages of different semantic categories for the two candidates.

	System	Trump, Oral (%)	Clinton, Oral (%)	Trump, Pres. debates (%)	Clinton, Pres. debates (%)	Trump, Speeches (%)	Clinton, Speeches (%)
Self	Diction	**5.3**	4.7	3.4	2.9	2.0	3.1
Human	Diction	9.1	7.1	**10.1**	9.1	7.9	9.1
Cognition	Diction	1.9	**2.7**	1.6	2.3	1.4	1.8
Adversity	Diction	0.9	0.7	0.9	0.9	**1.5**	0.8
Symbolism	Diction	1.4	1.2	1.5	1.4	**2.4**	1.5
Tenacity	Diction	5.3	4.7	3.4	2.9	7.3	**7.4**
Exclusive	LIWC	**4.3**	2.8	3.7	2.7	2.5	2.7
Affect	LIWC	5.6	5.0	5.3	5.0	**6.1**	5.5
Posemo	LIWC	3.6	3.5	2.9	3.3	3.5	**3.8**
Negemo	LIWC	1.9	1.4	2.4	1.7	**2.6**	1.6

find words corresponding to 'sacred terms' in the United States (e.g. *American*, *people*, *peace*, *rights*, etc.) while the *Adversity* category regroups words related to dangerous events (*crime*, *attack*, *crisis*, etc.), and the *Cognition* list contains terms such as *think*, *believe*, *know*, etc. Some dictionaries could be rather short as, for example, *Self* containing the pronoun *I* and some related forms (e.g. *me, my, I'd, I'm*).

For each analysed text, the system counts the percentage of words or expressions occurring in each category. For the two nominees and the three corpora, Table 6 depicts the percentage of occurrences of selected dictionaries (the origin is provided in the second column). This data confirms the high occurrence frequency of *Self* terms in Trump's dialect. In his written messages, this percentage is decreased significantly confirming our previous claim specifying that there is a real stylistic and rhetoric difference between Trump's oral and written communication.

The *Human* dictionary contains words related to persons (e.g. *family*, *friend*, …), family members (e.g. *father*, *child*) as well as personal pronouns (e.g. *we*, *our*, *they*, *she*). This last component explains the high frequency of this category, especially in Trump's dialect. For Clinton, the three corpora indicate similar percentages.

Clinton's rhetoric employs cognitive terms (*think*, *know*) more often leading to more nuanced statements. With the *Adversity* and *Symbolism* categories, one can see that the percentages of difference are small when considering the oral form between the two candidates. In the written form, Trump employs more hardship terms (1.5%) in his speeches compared to Clinton's usage (0.8%).

The *Tenacity* category regroups mainly auxiliary verb forms (e.g. *is, was, has, must, do*) and corresponds to an indicator of the speaker's persistence. Table 6 shows that such forms emerge more frequently in the written form while, in the oral genre, Trump uses them more often. Under the label *Exclusive*, one can find terms indicating an exclusion, an exception such as *but, without, rather, not, either*. This category appears more frequently with Trump in the oral genre (4.3 versus 2.8% for Clinton).

During an electoral campaign, emotions play an important role. For Bligh et al. (2010), the leader's charisma can be improved by showing and talking with positive emotions, by sharing an inspiration and announcing a bright future. In Table 6, the last three rows correspond more to this aspect with the *Affect* category which is simply the union of the positive emotion words (*Posemo* with *love*, *hope*, *help*, etc.) and negative emotion list (*Negemo* with *hate*, *fear*, *sad*, *war*, etc.). In this perspective, Trump uses emotional words or expressions more often. The data shown in Table 6 indicate that such terms tend to cover between 5 and 6% of the speeches, but only slightly more for Trump. The difference emerges when discriminating between the positive (*Posemo*) and negative (*Negemo*) emotion words. The latter occurs more frequently in Trump's speeches while the positive ones occur slightly more often with Clinton.

6. Conclusion

The political scientists Caesar, Thurow, Tulis, and Bessette (1981) postulate that 'speaking is governing'. Speaking is also an essential activity during an electoral campaign. Each candidate adopts his/her own communication strategy to present his/her programme, to convince the citizens of their leadership, and to motivate his/her sympathisers. The style and rhetoric are therefore essential for reaching these objectives. To detect and analyse those differences between Trump and Clinton, we have examined both the oral communication form (based on TV debates and interview transcripts) and the written form (speeches).

It is known that the oral and written genres present differences (Biber et al., 2002) as, for example, a dialogue implies more occurrences of verbs and pronouns. The current study shows some of them but it also indicates that Trump presents two distinct styles when analysing his oral and written form. This distinction does not appear with Clinton. Based on a graph visualizing the stylistic affinities and differences, one can see that Clinton's oral form, TV presidential debates, and speeches are relatively close. For Trump, a clear difference appears between the oral interventions (interviews and TV debates) and his speeches (certainly written by ghostwriters without Trump's close supervision).

Various global stylistic measurements demonstrate that Trump's oral style is direct, based on brief sentences composed of short words, with similar expressions reused many times. This choice can be explained by the determination to be understood by everybody. Clinton prefers uttering longer sentences with a richer vocabulary. She also tries to cover more topics. According to their grammatical category distribution, Trump's oral form is slightly more oriented towards action with more verb phrases (verbs and adverbs) while Clinton opts for a more descriptive rhetoric (more nouns, adjectives, prepositions, and determiners).

As a general conclusion, the style and rhetoric selected by Trump lets him appear as a strong masculine figure, talking like every American with energy, easy to understand with a nationalism colour (more *Symbolism* terms). His victory against the elites can also be explained by other factors such as the uprising of the people against Washington, the parties, the politicians, the news media, Hollywood, academia, etc. (Fisher, 2016) as well as some failures of the incumbent president and administration (Lichtman, 2016), and FBI investigations devastating Hillary's image. Finally, some explanations given by some of Trump's staffers can provide another light on Trump's style and rhetoric:

> She [Hillary Clinton] was defined as someone that people don't like and don't trust, and all we had to do was reinforce the existing narrative. (Balz & Rucher, 2016)

Disclosure Statement

No potential conflict of interest was reported by the author.

References

Arnold, E., & Labbé, D. (2015). Vote for me. Don't vote for the other one. *Journal of World Languages, 2*, 32–49.

Baayen, R. H. (2008). *Analysis linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Balz, D., & Rucher, P. (2016). How Donald Trump won: The insiders tell their story? *Washington Post*, November 9th.

Biber, Douglas, & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Biber, G., Conrad, S., & Leech, G. (2002). *Longman student grammar of spoken and written English*. London: Longman.

Bligh, M., Merolla, J., Schroedel, J. R., & Gonzalez, R. (2010). Finding her voice: Hillary Clinton rhetoric in the 2008 presidential campaign. *Woman's Studies, 39*, 823–850.

Boller, P. F., Jr (2004). *Presidential campaigns. from George Washington to George W. Bush*. Oxford: Oxford University Press.

Boyd, D. (2016). Reality check: I blame the media. *Points*, November 9th. Retrieved from https://points.datasociety.net/reality-check-de447f2131a3#.cosi2pmkr

Caesar, J. W., Thurow, G. E., Tulis, J., & Bessette, J. M. (1981). The rise of rhetorical presidency. *Presidential Studies Quarterly, 11*, 158–171.

Covington, M. A., & McFall, J. D. (2010). Cutting the Goridan knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics, 17*, 94–100.

Fisher, M. (2016). How Donald Trump broke the old rules of politics and won the White House. *Washington Post*, November 9th.

Francis, W. N., & Kucera, H. (1982). *Frequency analysis of English usage*. Boston, MA: Houghton Mifflin Co.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*, 267–297.

Hart, R. P. (1984). *Verbal style and the presidency. A computer-based analysis*. New York, NY: Academic Press.

Hart, R. P., Childers, J. P., & Lind, C. J. (2013). *Political tone. How leaders talk and Why*. Chicago, IL: The University of Chicago Press.
Kubát, M., & Cech, R. (2016). Quantitative analysis of US presidential inaugural addresses. *Glottometrics, 34*, 14–27.
Labbé, D. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics, 14*, 33–80.
Labbé, D., & Monière, D. (2003). *Le discours gouvernemental. Canada, Québec, France (1945-2000)* [The governmental discourse. Canada, Québec, France (1945–2000)]. Paris: Honoré Champion.
Labbé, D., & Monière, D. (2008a). *Les mots qui nous gouvernent. Le discours des premiers ministres québécois: 1960-2005*. Montréal: Monière-Wollank.
Labbé, D., & Monière, D. (2008). *Je est-il un autre ? Proceedings JADT, 2008*, 647–656.
Labbé, D., & Monière, D. (2013). *La campagne présidentielle de 2012. Votez pour moi!* Paris: L'Harmattan.
Lakoff, G., & Wehling, E. (2012). *The little blue book: The essential guide to thinking and talking democratic*. New York, NY: Free Press.
Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review, 97*, 311–331.
Lichtman, A. J. (2016). Predicting 2016: How the '13 keys to the White House' will turn in 2016. *HubPages*, June 7th. Retrieved from hubpages.com/politics/Predicting-the-2016-Election-the-13-Keys-to-the-White-House
Millbank, D. (2016). Trump's fake-news presidency. *Washington Post*, November 18th.
Mitchell, D. (2015). Type-token models: A comparative study. *Journal of Quantitative Linguistics, 22*(1), 1–21.
Muller, C. (1992). *Principes et Méthodes de Statistique Lexicale*. Paris: Honoré Champion.
O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. Proceedings 4th International AAAI Conference on Weblogs and Social Media, 122–129.
Paradis, E. (2011). *Analysis of phylogenetics and evolution with R*. New York, NY: Springer.
Pauli, F., & Tuzzi, A. (2009). The end of year addresses of the presidents of the Italian republic (1948–2006): Discourse similarities and differences. *Glottometrics, 18*, 40–51.
Pennebaker, J. W. (2011). *The secret life of pronouns. What our words say about us*. New York, NY: Bloomsbury Press.
Popescu, I. I. (2009). *Word frequency studies*. Berlin: Mouton de Gruyter.
Sainato, M. (2016). Email reveals Clinton camp spied on Sanders delegates before convention. *Observer*, November 14th. Retrieved from https://observer.com/2016/11/email-reveals-clinton-camp-spied-on-sanders-delegates-before-convention/
Savoy, J. (2015). Text clustering: An application with the *State of the Union* addresses. *Journal of the American Society for Information Science and Technology, 66*, 1645–1654.
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*, 63–75.
Sylwester, K., & Purver, M. (2015). Twitter language use to reflects psychological differences between Democrats and Republicans. *PLoS One, 10*(9), 1–18.
Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*, 24–54.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclid dependency network. In Proceedings of NAACL 2003, ACL, 252–259.
Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., … Patwardhan, S. (2005). OpinionFinder: A system for subjectivity analysis. Proceedings Empirical Methods for Natural Language Processing, Vancouver (BC), 34–35.
Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication, 29*, 205–231.
Yourish, K. (2016). Clinton and Trump have a terrible approval rating. Does it matter? *New York Times*, June 3rd.
Yu, B. (2008). Classifying party affiliation from political speech. *Journal of Information Technology & Politics, 5*, 33–48.
Yu, B. (2013). Language and gender in congressional speech. *Literary and Linguistic Computing, 29*, 118–132.