Analysis of the style and the rhetoric of the 2016 US presidential primaries

Jacques Savoy
University of Neuchatel, Switzerland

Abstract

This present article examines the verbal style and rhetoric of the candidates of the 2016 US presidential primary elections. To achieve this objective, this study analyzes the oral communication forms used by the candidates during the TV debates. When considering the most frequent lemmas, the candidates can be split into two groups, one using more frequently the pronoun 'I', and the second favoring more the 'we' (which corresponds to candidates leaving the presidential run sooner). According to several overall stylistic indicators, candidate Trump clearly adopted a simple and direct communication style, avoiding complex formulation and vocabulary. From a topical perspective, our analysis generates a map showing the affinities between candidates. This investigation results in the presence of three distinct groups of candidates, the first one with the Democrats (Clinton, O'Malley, and Sanders), the second with three Republicans (Bush, Cruz, Rubio), and the last with the duo Trump and Kasich, with, at a small distance, Paul. The over-used terms and typical sentences associated with each candidate reveal their specific topics such as 'simple flat tax' for Cruz, 'balanced budget' for Kasich, negativity with Trump, or critiques against large corporations and Wall Street for Sanders.

Correspondence:
Jacques Savoy, University of Neuchatel, rue Emile Argand 11, 2000 Neuchatel, Switzerland.
E-mail:
Jacques.Savoy@unine.ch

1 Introduction

The 2016 US presidential election was characterized by two figures, both unloved by the majority of Americans (Yourish, 2016). Donald Trump seemed sincere, authentic, saying what he thinks, putting aside political correctness. For him, all appearances and comments on the media were an opportunity for self-promotion. He believed that the repetition of a simple message, even if false (Millbank, 2016), is enough to persuade the citizens that it is true. His image was centered 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 pertinent interest, and without a real future. But Trump was able to beat all his opponents and won the nomination for the Republican party (21 July 2016).

Nominated by the Democrats (28 July 2016), Hillary Clinton always appeared as a cold woman, a member of the political establishment rejected by many people. She did not like the press and the media and, in return, they do not like her much either. This aspect could be related to her first years at the White House as an overqualified First Lady who wanted to play a principal role in politics (e.g. health care reform in 1993). For some people, she was even a crook and a liar, or, at least, dishonest (Sainato, 2016). When her campaign starts (14 April 2015), everything seemed simple and the road to the nomination seems without

doi:10.1093/llc/fqx007

any real problem. The presence of Bernie Sanders occupying a position more on the left demonstrates 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, especially during the general election campaign.

Even if Hillary's candidature appeared more natural, she needed to convince the Democrats and their sympathizers that she was the right person who can win the general election. Inside the Republican party, the outcome of the fight was more uncertain, in part by the larger number of candidates (seventeen vs. six), and the leading position occupied by Jeb Bush in the beginning of the primaries. Despite the now-known election outcome, the candidates' use of language during the primary season raises some pertinent research questions. How were the respective nominees able to win the primaries according to their speeches? Does the analysis of TV debates make it possible to detect their style and rhetoric? Can one discover the rhetorical features that can explain a Trump or Clinton success? Can one measure the stylistic distance between the candidates in both parties?

To answer these questions, the current study will focus on the US primaries' TV debates. Here, rhetoric is defined 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 for mainly esthetical value (Biber and Conrad, 2009). The analysis will use the oral communication form, a more direct and spontaneous way of interacting, reflecting more closely the personal style of each candidate. The style of the written messages (evident in prepared statements by the candidates) differs from the oral dialogue (Biber *et al.*, 2002). Moreover, the statements are certainly authored, at least in part, by a team of speechwriters. Therefore, the two forms of communication must be analyzed separately.

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

2 Related Work

Political texts have been the subject of various studies discussing different aspects of them. Focusing on governmental speeches written in French, 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 statements of French governments (Labbé and Monière, 2003; 2008a) as well as a collection of press releases covering the French presidential campaign of 2012 (Labbé and Monière, 2013; Arnold and Labbé, 2015). Similar research has been conducted with other languages, such as Italian (Pauli and Tuzzi, 2009). From these analyses, one can observe, for example, that governmental institutions tend to smooth out the differences between political parties when exercising command. Moreover, the temporal period of the documents constitutes an important factor explaining the variations between presidents or prime ministers. The presence of a strong leader is usually accompanied with a real change in the style and vocabulary of governmental speeches (Labbé and Monière, 2003; Savoy, 2015c).

Focusing on the USA, recent studies confirm these findings as, for example, using the 'State of the Union' (Savoy, 2015a) or inaugural addresses (Kubát and Cech, 2016). Beside time frame, exceptional events (e.g. worldwide war, deep economic depression) may change noticeably the style. These results present also the stylistic evolution over more than two centuries and can be compared to those achieved using traditional methods as, for example, by Lim (2002).

Differentiations between political parties can however be observed as, for example, studies based on tweets (Sylwester and 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 using a training corpus, Laver *et al.* (2003) describe a methodology to extract 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 (political) feminine figures tend to use emotional words more frequently and employ more personal pronouns than men. A more general overview of using different computer-based strategy to detect and extract topical information from political texts can be found in (Grimmer and Stewart, 2013).

The web-based communication (e.g. tweets, blogs, chats) was used by O'Connor *et al.* (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 because 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).

Also grounded on several dictionaries (or categories), Young and Soroka (2012) describe how one can detect and measure sentiments appearing in political texts. The suggested approach is rather similar to O'Connor *et al.*'s work (2010), counting the frequency of occurrence of words appearing in a dictionary of positive or negative emotion words. Using also some lists of words, Hart (1984) has designed and implemented a political text analyzer called DICTION. Based on US presidential speeches, that study presents the rhetorical and stylistic differences between the US presidents from Truman to Reagan, while a more recent book (Hart *et al.*, 2013) exposes the stylistic variations from G. W. Bush to Obama. Using the DICTION system, Bligh *et al.* (2010) analyze 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, Linguistic Inquiry and Word Count (LIWC; Tausczik and 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. grounded 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 *et al.* (2007) were able to determine the personalities of different political candidates (US presidential election in 2004). They defined the psychological portrait both on single measurements (e.g. the relative frequency of different pronouns, positive emotions) and using a set of composite indices reflecting the cognitive complexity, presidentially, or honesty of each candidate. These personality measurements were in agreement with different opinion polls. For example, G. W. Bush used more frequently the pronoun 'I', positive emotion words (e.g. happy, truly, win), and the future tense. The public perceived J. Kerry as a kind of depressed person, serious, somber, and cold, adopting more frequently negative emotion expressions (e.g. sad, worthless, cut, lost) and physical words (e.g. head, ache, sleep).

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

3 Electoral Corpus

To analyze the rhetoric and style adopted by the candidates during the primaries of the 2016 US primary election, the transcripts of the TV debates have been downloaded from the Internet (mainly from the Web site www.presidency.ucsb.edu). For the Republican candidates (Jeb Bush, Ted Cruz, John Kasich, Rand Paul, Marco Rubio, and Donald Trump), twelve TV debates were organized, from the 1st one held on 6 August 2015 with ten candidates, to the last one organized on 10 March 2016 with four candidates. For the Democrats (Hillary Clinton, Martin O'Malley, and Bernie Sanders), one can count nine TV debates held from 13 October 2015 (with five candidates) to 9 March 2016 (with two candidates).

Due to space limitations, not all possible candidates have been retained. Some persons never appear in a TV debate (e.g. Pataki (R), Jindal (R), Lessig (D)) or just appear in one (e.g. Webb (D)) or two debates (e.g. Walker (R)) while others have been ignored because they have played a minor role during the electoral campaign (e.g. Carson (R) or Christie (R)).

From a stylistic point of view, this corpus is homogenous, corresponding to an oral communication form, extracted from a short period of time, and with the same main objectives (convincing the people, answering questions, presenting their ideas and solutions). Several factors influencing the style are therefore fixed. The remaining variations can be largely explained by the speaker.

Even if the topics are not directly and fully controlled by the candidates, the debate format corresponds to a more spontaneous form of communication, able to reveal more closely the real person behind her/his projected image. Of course, one can raise the question of the speaker's spontaneity because Donna Brazile, who worked for CNN, provided, at least once, prepared questions to Clinton before a debate. This phenomenon is assumed to be the exception rather than the norm.

One can consider that electoral speeches delivered by the candidates correspond also to an oral communication form and thus can be included in our corpus. 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 vs. 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.

Table 1 reports the vocabulary size (number of distinct word types) for each candidate (under the row 'Voc.') as well as the total number of word tokens (row 'Token'). To understand the difference between word type and 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' each occur twice. The set of all distinct word types forms the vocabulary, denoted by V, while the text length is represented by n.

As shown in Table 1, Paul and O'Malley correspond to the smallest values, both being present for a relatively short period of time during this electoral campaign. Clinton appears with the largest number of tokens, followed by Sanders, Trump, Rubio, and Cruz.

4 Evaluation of Stylistic Characteristics of the Candidates

To highlight 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 ones is certainly a good starting point, as shown in the first sub-section. The second proposes to consider four overall stylistic measurements and applies them to the different candidates while the last sub-section describes the differences in the distribution of the grammatical categories between candidates.

Table 1 Some statistics about candidates' speeches and comments

Measurements	Bush	Cruz	Kasich	Paul	Rubio	Trump	Clinton	O'Malley	Sanders
Voc.	2,367	3,373	2,539	1,542	3,203	2,614	3,537	2,019	2,954
Token	23,207	37,459	33,807	11,856	44,415	49,718	61,991	15,824	53,172

Table 2 The top ten most frequent lemmas according to TV debates

Bush	Cruz	Kasich	Paul	Rubio	Trump	Clinton	O'Malley	Sanders
be	be	the	be	be	be	be	**we**	be
the	the	be	the	the	**I**	the	the	the
to	and	**we**	to	to	the	to	be	**I**
we	to	to	**I**	and	**we**	**I**	and	to
that	**I**	and	**we**	that	and	and	to	and
and	that	**I**	an	**we**	to	**we**	of	of
an	**we**	an	and	an	have	that	an	that
I	of	have	that	**I**	an	have	that	**we**
of	an	in	have	of	**you**	of	**I**	an
have	**you**	**you**	in	in	**it**	an	in	in

Note. The personal pronouns are depicted in bold.

4.1 Most frequent lemmas

To analyze the rhetoric and style of presidential writings, the first quantitative linguistics studies focused on the word usages and their frequencies. As the English language has a relatively simple morphology, working on inflected forms (e.g. 'we, us, ours', or 'wars, war') or lemmas (dictionary entries such as 'we' or 'war' from the previous example) often lead to similar conclusions.

To define the lemma of each token, the part-of-speech (POS) tagger developed by Toutanova *et al.* (2003) was first applied. Given a sentence as input, this system is able to add the corresponding POS tag to each token. For example, from the sentence 'But I also know this problem is not going away', the POS tagger returns 'But/CC I/PRP also/RB know/VBP this/DT problem/NN is/VBZ not/RB going/VBG away/RB ./.'. Tags may be attached to nouns (NN, noun, singular, NNS noun, plural, NNP proper noun, singular), verbs (VB, lemma, VBG gerund or present participle, VBP non-3rd-person singular present, VBZ 3rd-person singular present), adjectives (JJ, JJR adjective in comparative form), personal pronouns (PRP), prepositions (IN), and adverbs (RB). These morphological tags (Marcus *et al.*, 1993) correspond mainly to those used in the Brown corpus (Francis and Kucera, 1982). With this information one can 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. detects/VBZ → detect/VB).

Our first analysis considers the most frequent lemmas occurring in the oral interventions of the candidates during the TV debates of the primaries. Unsurprisingly, the article 'the' and the verb 'be' (lemma of the word types am, is, are, was, etc.) appear regularly in the first two ranks. Looking at the most frequent lemmas in the Brown corpus (Francis and Kucera, 1982), the first two are the same, but after them the order changes. In the Brown corpus, the top ten most frequent lemmas are as follows: the, be, of, and, to, a, in, he, have, it.

Table 2 reports the top ten most frequent lemmas for each selected candidate. 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 (but does not appear in the top ten positions in the Brown corpus), and to a lesser extent 'you' or 'it'. Even if pronouns are more frequent in dialogue or in oral form than in written communication (Biber and Conrad, 2009), the high frequency of 'I' and 'we' is a fundamental

Table 3 Four global stylistic measurements according to TV debates

Measurements	Bush	Cruz	Kasich	Paul	Rubio	Trump	Clinton	O'Malley	Sanders
MSL	17.3	19.4	18.3	17.2	18.7	**13.7**	20.5	**21.9**	19.7
LD (%)	40.7	**44.6**	38.4	40.3	39.2	**36.6**	40.4	43.1	43.6
BW (%)	24.4	**26.4**	20.8	21.9	23.8	**18.3**	24.1	25.6	**26.4**
TTR	36.1	37.3	33.9	34.0	33.3	**29.7**	36.2	**37.9**	36.1

Note. Extreme values are depicted in bold.

characteristic of political speech. Moreover, usually the 'we' is more associated with the government, the president, or the prime minister (Pennebaker, 2011). In 'State of the Union' speeches by Obama, the absolute frequency of the article 'the' and the pronoun 'we' is very similar.

In this table, one can see another interesting fact related to the frequencies of the pronouns 'we' and 'I'. Former governors tend to use more frequently the 'we' than the 'I' (e.g. Bush, Kasich) with O'Malley having a very distinctive style in this point of view. Usually Senators (e.g. Cruz, Paul, Clinton, Sanders) tend to prefer using the pronoun 'I', at least during an electoral campaign. The candidates who stayed longer in this campaign have a clear preference for the 'I' over the 'we'. The pronoun 'we' stays ambiguous (Who is behind the 'we'? Myself and the future government? Me and the people? Me and the workers? Me and the Congress?). Finally, the champion in the usage of 'I' is Trump who clearly has adopted a distinct style in the campaign, putting the light more on his ego.

4.2 Global stylistic measurements

To define an overall measurement of the style, various studies have proposed different measures. As a first indicator, one can consider the mean sentence length (MSL) reflecting a syntactical choice. The sentence boundaries are defined by the POS tagger (Toutanova *et al.*, 2003) and correspond to 'strong' punctuation symbols (namely, periods, question, and exclamation marks). Usually, a longer sentence is more complex to understand, especially in the oral communication form. Using the 'State of the Union' addresses given by the Founding Fathers, this average value is 39.6 (with Madison depicting the highest MSL with 44.8 tokens/sentence). With Obama, the MSL decreases to 18.5 tokens/sentence. These examples indicate clearly that the style is changing over time. Currently, the preference goes to a shorter formulation, simpler to understand for the audience.

During the 2016 primaries, the average value per candidate reported in Table 3 (label MSL) confirms this tendency with a mean varying from 21.9 (O'Malley) to 13.7 (Trump). One can also see that all values are relatively close to 19, except for Trump who is adopting an even more simple and direct communication style. The presence of long sentences (O'Malley, Clinton, Sanders) indicates a substantiated reasoning or specifies the presence of detailed explanations. Even if a long sentence is required, its length does not guarantee an easy understanding.

As a second stylistic indicator, the lexical density (denoted LD in Table 3) can be used to reveal the informativeness of a text (Biber *et al.*, 2002; Hewings *et al.*, 2005). The formulation is shown in Equation (1) where the variable $n(t)$ indicates the total number of tokens (or the text length) of a text t, *function words*(t) the number of function words in t, *lexical word*(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 all other grammatical categories, namely, determiners (e.g. the, this), pronouns (e.g. you, us), prepositions (e.g. to, in), conjunctions (e.g. and, or), modal verbs, and auxiliary verb forms (e.g. has, would, can). The list of functional words for the English language contains 409 entries (extracted from the LWIC dictionaries; Tausczik and Pennebaker, 2010). As depicted in Equation (1), this LD value is given in percentage over the number of tokens.

$$\mathrm{LD}(t) = {}^{lexicalword(t)}\!/_{n(t)} = 1 - {}^{functionword(t)}\!/_{n(t)} \quad (1)$$

A relatively high LD percentage indicates a more complex text, containing more information. Using the transcripts of the TV debates, the LD values vary from 36.6% (Trump) to 44.6% (Cruz). Trump's style appears, here too, as distinct from the others, providing his answers and comments around functional words. Cruz adopts an opposite style, focusing more on topical forms and expressions.

As an additional global stylistic measurement, the frequency of big words (composed of six letters or more, and denoted BW) can be analyzed (Tausczik and Pennebaker, 2010). A text or a dialogue with a high percentage of BW tends to be more complex to understand. This fact is confirmed by recent studies:

> '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.' (Lakoff and Wehling, 2012, p. 41)

This rhetoric problem was recognized by previous US president such as President Johnson who told his speechwriters: 'I want four-letter words, and I want four sentences to the paragraph.' (Hart, 1984). Table 3 indicates that the percentage of BW varies from 18.3% (Trump) to 26.4% (Cruz, and Sanders). Grounded on the MSL, LD, and BW indicators, one can see that Trump is adopting a more direct communication style, selecting simple words and producing short sentences. Senators Cruz or O'Malley have a more sophisticated communication style, employing longer sentences and a more complex lexicon.

The Type-Token Ratio (TTR) or the relationship between the vocabulary size and the number of word types (Baayen, 2008) corresponds to our last global stylistic measure. High values indicate the presence of a rich vocabulary showing that the underlying text exposes many different topics or that the author tends to present a theme from several angles with different formulations. To compute this value, one can divide the vocabulary size (number of types) by the text length (number of tokens). This estimator has the drawback of being instable, tending to decrease with text length (Baayen, 2008). To avoid this problem, a better computation is provided in (Covington and McFall, 2010) or (Popescu, 2009), suggesting taking the moving average of TTR. This computation technique has been adopted.

From data depicted in Table 3, one can see that the TTR value reaches a minimum of 29.7 (Trump) to a maximum of 37.9 (O'Malley). This value indicates that Trump prefers to reuse the same words and expressions, repeating his main ideas and convictions. On the other hand, O'Malley or Cruz (TTR: 37.3) have opted for a larger coverage requiring a larger number of distinct words and phrases. It should be noted that, regarding the four indices, Clinton is closer to O'Malley or Cruz than to Trump.

4.3 POS distribution

The analysis of the style can be grounded on the relative frequencies of the different POS or grammatical categories. Table 4 presents these distributions, in percentage, over the nine candidates. Two main syntactic constructions can be selected by the speaker, namely, using more frequently verb phrases (composed of verbs and adverbs), or choosing more often noun phrases (with nouns, adjectives, determiners, and prepositions). In addition, Table 4 reports the percentage of pronouns, names, conjunctions, and others forms (e.g. number, inserts). In this table, the maximum value per grammatical category is shown in bold, and the minimum in italics.

Data depicted in Table 4 indicate that O'Malley is the candidate choosing most frequently the noun construction (largest percentage of nouns, and adjectives). The verb phrase is used most frequently by Trump with the largest percentage of verbs and adverbs and the lowest frequencies of nous, determiners, and prepositions. The difference between these two candidates is characteristic, with O'Malley more oriented toward an explanation requiring usually more nouns while Trump turns toward the action and its high usage of verbs. Table 4 confirms that Trump owns a style very distinct than the others. Moreover, Trump is using pronouns less—except I—than the others.

To obtain an overall measure of the intensity of the action over the descriptive part of a text, Kubát

Table 4 POS distribution according to TV debates

Measurements	Bush (%)	Cruz (%)	Kasich (%)	Paul (%)	Rubio (%)	Trump (%)	Clinton (%)	O'Malley (%)	Sanders (%)
Noun	18.8	19.1	17.8	17.8	18.1	*15.9*	17.7	**20.6**	19.9
Name	5.4	**8.4**	4.9	5.6	5.5	*4.6*	5.2	5.1	6.3
Pronoun	1.9	2.3	2.5	2.1	2.5	*1.9*	2.5	**3.0**	2.4
Adjective	6.9	6.5	*5.4*	6.5	6.1	6.3	6.6	**7.6**	7.4
Verb	25.7	24.8	27.2	27.8	25.4	**30.2**	26.7	*21.1*	23.6
Adverb	6.7	*6.1*	7.7	7.7	8.1	**10.2**	7.8	7.0	7.1
Determiner	**13.1**	11.4	11.6	11.9	12.5	10.7	*10.6*	11.6	11.5
Preposition	16.0	14.5	16.2	15.2	16.0	*13.8*	**17.2**	17.2	15.6
Conjunction	*3.7*	4.7	**5.0**	3.9	4.2	4.4	4.3	4.8	4.0
Other	1.8	2.1	1.8	1.6	1.6	2.0	1.4	2.0	2.0
Q-index	78.8	79.2	**83.5**	81.1	80.7	82.6	80.1	*73.4*	76.0

Note. The maximum value per grammatical category is shown in bold, and the minimum in italics.

and Cech (2016) suggest to compute the ratio between the proportion of verbs divided by the sum of the proportion of the verbs and adjectives. The underlying idea is to quantify the activity by verbs while the descriptiveness of a text is represented by the proportion of adjectives. The last row of Table 4 reports the values of this Q-index for all candidates. Kasich depicts the highest value leading clearly more toward action. With a similar value and following the same tendency, one can find Trump and Paul. On the other hand, O'Malley shows the smallest Q-index value indicating more a text oriented toward description. The same feature can be assigned to Sanders.

5 Evaluation of Topical Characteristics of the Candidates

The previous section focusses mainly on stylistic features, both at the lexical and syntactical level. When looking more at the content of their utterances, one can also observe differences between the candidates. This analysis is based on the thematic concentration of a text, providing a first overview at the recurrent topical terms used by each candidate. In the second sub-section, an intertextual distance is presented and used to derive graphs representing stylistic and topical affinities between the candidates. Finally, the terms and sentences specific to each candidate will be computed and some examples will be given.

5.1 Thematic concentration

Recently, Popescu (2009) and Cech *et al.* (2015) have proposed an *h*-point to measure the thematic concentration of a text. To compute this value, the word types are ranked according to their absolute frequency, from the most frequent to the least frequent one. The *h*-point is defined as the point where the frequency is equal to the rank. From this *h*-point, one can assume that types appearing before the *h*th rank are function words while those occurring after correspond to lexical or topical words (the very vocabulary).

The frequency table reveals that some lexical terms can appear before the *h*-point. Those terms correspond to recurrent thematic words or expressions on which the author wants to insist. Table 5 reports both the *h*-point and the first four recurrent topical words. For all candidates, except for Cruz or Paul, the term 'people' appears in the top four most frequent topical words. As another frequently used word, one can see 'say' (five candidates), 'country, know, go' (four candidates), and 'president, think' for three candidates. This set of words indicates the political electoral parlance or recurrent words used by politicians during an election. One can compare them to those employed by US presidents in their inaugural addresses (Kubát and Cech, 2016). In common, one can find the terms 'people', or 'country'. Presidents are also using very frequently the words 'government', 'world', and the adjectives 'free', 'great', or 'new'. None of these terms appears here. In this campaign, none of the adjectives is very

Table 5 *h*-point and most frequently used thematic words per candidate

Measurements	Bush	Cruz	Kasich	Paul	Rubio	Trump	Clinton	O'Malley	Sanders
h-point	52	47.8	66.75	*39*	76	77.3	84.5	43.5	**85**
PTC (%)	4.99	6.19	8.49	4.82	8.04	9.89	9.42	*4.26*	**12.12**
	need	know	people	think	people	people	think	need	people
	people	Donald	know	say	go	go	people	people	think
	country	say	go	want	president	say	know	country	country
	president	president	want	go	country	know	say	make	say

Note. The maximum value is shown in bold, and the minimum in italics.

frequently used, but more verbs can be observed (e.g. say, know, go, want), and the noun 'president'.

Finally, Table 5 shows that with shorter available text as for Paul or O'Malley, the corresponding *h*-point is lower than for the others. In fact, the *h*-point tends to increase with the text length (denoted *n*) as one can see from Equation (2) in which *a* is a constant.

$$n = a \cdot h^2 \text{ or } h = \sqrt{n/a} \tag{2}$$

To have a better overall measurement of the topical concentration of a text, Cech *et al.* (2015) define a proportional thematic concentration (PTC) defined as:

$$PTC = {}^{1}/_{n_h} \sum_{r'<h} f(r') \tag{3}$$

in which n_h indicates the frequency of all word types before the *h*-point, and f(*r'*) the frequency of 'lexical' words appearing before the *h*-point.

According to this formulation, when all word types appearing before the *h*-point are functional words, the PTC value is 0. On the other hand, when all those types are lexical words, PTC reaches the maximum value of 1.0 or 100%. In Table 5, the second row indicates the PTC values for all candidates, showing clearly that Sanders presents the highest PTC value (12.12%) while O'Malley exposes the smallest (4.26%). Sanders' answers and remarks are clearly more focused on a few topics. With Trump and Clinton, one can find also relatively high PTC values compared to the other candidates. Both are preferring to repeat their arguments instead of introducing other subjects. On the other hand, Bush, Paul, and, to a lesser extent, Cruz are closer to O'Malley's PTC value, being able to present alternative formulations or covering more distinct subjects.

5.2 Stylistic and topical distance between candidates

As each candidate is represented by her/his remarks in the TV debates, one can compute a distance reflecting their similarities (Labbé and Labbé, 2006). A text is, however, a composite item in which one finds both the style with its lexical, syntactical, or discourse factors, and the recurrent words belonging to the topics. To distinguish between these two main components, the first map will use the stylistic aspects while the second will take into account the topical elements. Splitting the vocabulary into two distinct parts is relatively known in stylistic (Damerau, 1975), authorship attribution (Argamon and Levitan, 2005; Stamatatos, 2009), or in quantitative linguistics studies (Tuzzi, 2010). The current analysis follows this principle to consider the style, on the one hand and, on the other, the content.

To reflect the style, one can consider the *k* top most frequent lemmas from our electoral corpus. No general theory specifies precisely the *k* value, but a value from 200 to 500 represents a pertinent choice justified by various studies (Savoy, 2015b). Taking another strategy, one can use the *h*-point splitting the vocabulary into two parts. As shown previously, the *h*-point is rather small, and thus, the words reflecting more the style (e.g. functional words) are rather limited. In the current study, the stylistic elements will be the words appearing in the functional words list (409 entries), and used previously in defining the LD measure.

The intertextual distance between Text A and Text B is computed according to Equation (4) (Labbé, 2007), in which V_A (or V_B) indicates the

vocabulary of Text A, tf_{iA} (respectively tf_{iB}) denotes the absolute term frequency of the ith word type in Text A, and n_A (respectively n_B) the length of Text A.

$$dist(A, B) = \sum_{i \in V_A \cup V_B} |tf_{iA} - tf_{iB}| \Big/ (n_A + n_B) \quad (4)$$

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 the largest text is modified as follows:

$$tf'_{iB} = tf_{iB} \cdot {}^{n_A}/_{n_B} \quad (5)$$

Other intertextual measures have been chosen as, for example, the chi-square (Grieve, 2007), the Delta (Burrows, 2002), or using the Kullback-Leibler divergence (Zhao and Zobel, 2007). The Labbé's measure (Labbé, 2007) has however demonstrated its effectiveness in various authorship attribution problems such as in literary works (Labbé and Labbé, 2006), in historical newspaper articles (Savoy, 2013), or in political speeches (Savoy, 2015d).

Grounded on this measure, one can compute the intertextual distance between all nine candidates. Displaying directly the 9 × 9 matrix containing these distances has a limited interest. Knowing that this matrix is symmetric and that the distance to itself is nil, we still have (81 − 9)/2 values. A better solution is to apply a clustering method (e.g. hierarchical clustering built on the complete link) to visualize the different groups of candidates according to their stylistic profiles. Recently, such distance matrices can be represented by a tree-based visualization method respecting 'approximately' the real distances between all nodes (Baayen, 2008). This new representation has been chosen, and the result is displayed in Fig. 1 obtained using the R software (Paradis, 2011; Saitou and Nei, 1987). Using graphical views to represent results of stylistic analysis is not new in quantitative linguistics studies. More often however, such displays correspond to

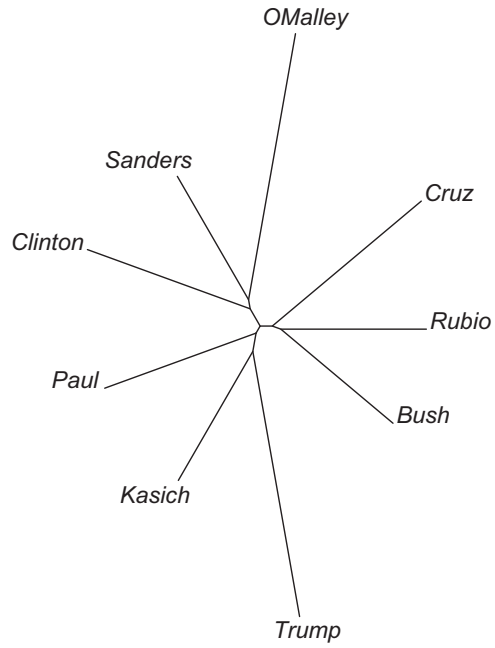

Fig. 1 Stylistic distance between candidates

scatterplots, principal component analysis views, or correspondence analysis figures (Lebart *et al.*, 1998; Greenacre, 2017). Tree-based graphs are more appropriate when displaying similarities between author profiles or stylistic affinities between works (Labbé and Labbé, 2006; Labbé, 2007).

In this figure, the distance between two candidates is indicated by the lengths of the lines connecting them. For example, starting with Bush, one can follow the branch until reaching the central point, then one can go along the lines leading to the second person (e.g. Trump or Clinton).

To generate Fig. 1, only the functional words (409 entries) have been used to reflect the style of each candidate. Based on this perspective, the longest distance (0.251) connects Trump and O'Malley, and the second longest (0.181) links Paul to O'Malley. The third longest distance (0.180) can be found between Trump and Cruz. The two closest candidates are Bush and Rubio (0.104), while the second shortest distance (0.114) joins Clinton with Sanders. More generally, Fig. 1 depicts three main groups, one

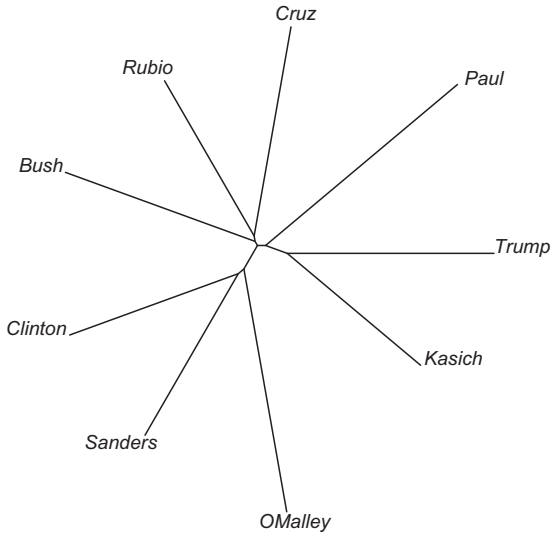

Fig. 2 Topical distance between candidates

Democrat, and two Republicans. From a stylistic point of view, the Republican cluster {Bush, Cruz, Rubio} is well separated from the second Republican group {Kasich, Trump, and with a smaller additional distance Paul} as well as for the Democrats {Clinton, O'Malley, Sanders}.

To build Fig. 2, the intertextual distance is computed according to topical words, and no word used to draw Fig. 1 is present in the elaboration of Fig. 2. To achieve this, the computation ignored all functional words for all texts. In this figure, the longest distance (0.510) connects Paul and O'Malley, and the second longest distance (0.496) links Trump to O'Malley. The third longest distance (0.467) can be found between Cruz and O'Malley. The two closest candidates are Clinton and Sanders (0.346), while the second shortest distance (0.360) joins Trump with Kasich.

As in Figs 1 and 2 reveals two Republican groups with the same members as in Fig. 1. Trump's topics are relatively similar to those presented by Kasich and, with some distance, with those exposed by Paul. With a longer distance, one can find the team Rubio and Cruz, with Bush having some affinities with this pair. On the bottom part, one can find the Democrats, with a shorter distance between Clinton and Sanders than with O'Malley.

5.3 Most specific terms

The analysis of the most frequent terms reveals some important themes of the 2016 primary election. But each candidate wants to promote his/her own specific point of view on some issues, and must underline his differences with others. Just considering the ten most frequent words, similar sets appearing with each candidate and the difference between them lies on their ranking as shown previously. Moreover, such analysis reveals more the style than the preferred themes. For example, from data depicted in Table 2, Trump, Clinton, and Sander prefers using 'I' instead of 'we', while Kasich, O'Malley, and Bush are using more frequently the pronoun 'we'.

Thus, which keywords or expressions can well describe each candidate and can be used to denote his/her difference? How can one define them, and, from a statistical point of view, be sure that the proposed terms are significantly over-used by the candidate? To measure the specificity attached to a term (Lafon, 1980; Muller, 1992), the corpus is split into two disjoint parts denoted P_0 and P_1. For a given term t_i, its absolute 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 by a given candidate, while P_1 denotes all other comments and remarks. Thus, for the entire corpus, the absolute 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 $n = n_0 + n_1$.

It is assumed that, for any term t_i, its distribution follows a binomial law, with parameters n_0 and $p(t_i)$ representing the probability of the term t_i being randomly selected from the entire corpus. Using the maximum likelihood principle, this probability is estimated as $p(t_i) = (tf_{i0}+tf_{i1})/n$. Of course, other models can be used as, for example, the hypergeometric one (Baayen, 2008) which could be viewed as the exact distribution. However, the binomial formulation is easier to use and gives a good approximation.

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

Table 6 The top ten most specific terms per candidate

Bush	Cruz	Kasich	Paul	Rubio	Trump	Clinton	O'Malley	Sanders
relates	Donald	Ohio	conservative	be	I	Senator	actually	Secretary
DC	flat	balanced	war	century	very	Sanders	Anderson	Street
signal	amnesty	budget	think	why	tremendous	to	Andrea	major
proven	note	we	Ferguson	someone	nobody	comprehensive	Maryland	Wall
serious	court	formula	Kentucky	issue	going	I	we	campaign
strategy	IRS	surplus	bowling	America	Mexico	income	ISIL	fossil
need	Islamic	secondly	scrutiny	they	not	try	Lester	fuel
caliphate	Washington	hole	borrow	this	Jeb	Republicans	freedom	class
brother	tax	Pentagon	fault	he	excuse	affordable	sort	billionaire
status	Texas	growth	Fed	operating	deal	more	nation	wealth

difference between these two values indicates a deviation from the expected behavior. To obtain a more precise definition of 'large', the binomial variance (defined as n_0 . p(t_i) . (1-p(t_i))) is used. Equation (6) 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.

$$Zscore(t_{i0}) = \frac{tf_{i0} - n_0 \cdot p(t_i)}{\sqrt{n_0 \cdot p(t_i) \cdot (1 - p(t_i))}} \qquad (6)$$

Applying this procedure, the term specificity can be computed according to the text P_0. Those 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 0). With a positive Z score larger than a fixed threshold δ (e.g. 3), one can conclude—with less than 1% chance of error—that the term is 'significantly over-used' in 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 -δ) indicates than the corresponding term is significantly under-used in P_0.

Table 6 shows the top ten most over-used terms for each candidate. One can see the presence of expressions related to the dialogue between candidates ('Senator Sanders' by Clinton, 'Donald' with Cruz, 'Jeb' under Trump). The relationship of some candidates to their origin is also represented ('Ohio' with Kasich, 'Texas' for Cruz, 'Kentucky' with Paul, 'Maryland' for O'Malley).

A more interesting finding is the presence of the pronoun 'I' in the most over-used terms by only two runners: Trump and Clinton. A candidate who wants to stay in the race must put forward him/her-self. After all the election is the process to select one person. Of course, behind this person, a political program must also appear. Some of the terms depicted in Table 6 give some indications about this aspect as, for example, 'IRS, tax, amnesty' with Cruz, 'fossil, fuel, Wall, Street' with Sanders. Similar themes appear under several candidates with different terminology such as 'caliphate' (Bush), 'Islamic' (Cruz), 'ISIL' (O'Malley). For Clinton, the term 'affordable' must be related to the Affordable Care Act (or health insurance reform).

5.4 Most specific sentences

Providing the most over-used terms is sometimes not enough to have a clear understanding of the candidate's position on a given issue. Can one be more precise than the simple sequence of isolated words such as 'balanced', 'budget', 'we' (Kasich) or 'IRS', 'tax' (Cruz)? One possible approach is to extract a reduced set of specific sentences from each candidate. Such a sentence can be defined as the one having the largest number of over-used terms. As it is extracted from a transcript, the sentence is not necessarily syntactically perfect.

Based on this definition, one can read some examples of the most specific sentences per candidate. As an interesting first case, one can analyze the most characteristic sentence from Kasich's comments, which is the following:

> 'I have *balanced budgets*, the federal *budget*, the state of *Ohio budget*, *we*'re running a 2 billion dollar *surplus*, *we*'re up 400,000 *jobs*, and in *Washington we* were able to have significant *job growth* whenever *we balanced* the *budget* of which I was the *architect*.' (J. Kasich, 6 February 2016)

In this example, terms having a Z score larger than 5.0 are depicted in italics. The sentence is longer that the MSL (18.3) for this candidate. From a lexical perspective, one can see the over-used term 'budget', usually more frequently used in governmental speeches than in the electoral ones. Here the candidate wants to put forward his competence in generating balanced budgets as Governor of Ohio. During an electoral campaign, the word 'tax' is clearly more recurrent than 'budget' to discuss financial issue, in part because the term 'tax' is closer to citizens' perception than the budget is. As another aspect, one can view that the preferred pronoun is 'we' and not 'I'. With this choice, the person can be viewed by the audience as distant and cold (Pennebaker, 2011). Looking back to Table 6, one can see that the terms 'Ohio', 'balanced', 'budget', and 'we' and the first four most specific words describing Kasich's utterances.

For Cruz, the most significant words depicted in Table 6 have a clearer meaning when reading two of his most specific sentences.

> 'So the way *you* do it is *you* pass a *tax* plan like the *tax* plan I've introduced: a *simple flat tax*, 10 percent for individuals, *and* a 16 percent business *flat tax*, *you abolish* the *IRS and* here's the critical point, *Maria*, the business *flat tax enables* us to *abolish* the corporate income *tax*, the death *tax*, the *Obamacare taxes*, the payroll *taxes*, *and* they're border-*adjustable*, so *every* export pays no *taxes whatsoever*.' (T. Cruz, 14 January 2016)

> '*And* I'*ll* tell *you*, Hugh *you* know, it's interesting now that *Donald* promises that he *will* appoint *justices who* . . . *who will defend religious liberty*, but this is a man *who*, for 40 years, has given money to *Jimmy Carter*, to Joe Biden, to Hillary Clinton, to *Chuck Schumer*, to *Harry Reid*.' (T. Cruz, 25 February 2016)

In this first example, the fiscal question appears with a proposition for a 'simple flat tax' and in the second, an attack against Trump, but a religious concern appears also in the background. These examples show that Cruz's rhetoric is more complex with the highest LD mean and the largest percentage of BW (see Table 3). Moreover, Cruz's explanations tend to include more names (see Table 4), and many of them (IRS, Maria, Obamacare, Donald, Carter, Biden, . . .) occur in these examples.

As demonstrated previously, Trump opted for a simple and direct communication style, preferring short sentences with simple words. The following remarks illustrate these aspects.

> 'They *don't* like seeing *bad trade deals*, they *don't* like seeing higher taxes, they *don't* like seeing a loss of their jobs where our jobs *have just been* devastated.' (D. Trump, 10 March 2016)

> '*I'm* spending all of *my* money, *I'm not* spending, *I'm not* getting any, *I* turned down, *I* turn down so much, *I* could *have* right now from special interests and donors, *I* could *have* double and triple what *he's* got.' (D. Trump, 16 September 2015)

> '*Just excuse me*, one second, Rand, . . . if *you don't* mind, Rand, *you* know, *you are* on last, *you do have your* 1 percent.' (D. Trump, 16 September 2015)

In a few words, Trump is able to talk not about a single but a few topics. This sentence was selected through the over-used words 'I', 'not', 'deal' (see Table 6), as well as 'bad', 'do', 'have', and 'just'. In the three sentences above, most of the words are less than six letters long (short words), and many of them are functional words (explaining his low LD). Moreover, these examples demonstrate the frequent use of symploces (repetition of the same formulation, e.g. they don't like seeing . . .) in Trump's comments. The last example illustrates how Trump can push away his opponents to place himself in the center of the debate.

From the set of the ten most over-used terms corresponding to H. Clinton depicted in Table 6, the first following sentence contains four (Senator, Sanders, to, I). As previously with Trump, the first person singular pronoun is clearly over-used (I, me). This phenomenon appears in other languages and countries when analyzing electoral speeches as, for example, in France (Labbé and Monière, 2008b). A presidential electoral process can be positioned around one person or around a few issues (or programs). In the current study, two candidates (Clinton and Trump) have opted for centering their communication around their person. In the following sentences, one can also see two topics, usually more related to the Democrats, namely education, and health care reform.

> 'Look, *I have* the greatest respect *for Senator Sanders* and *for* his *supports* and *I*'m going *to* keep working as *hard* as *I* can *to* reach as many people of all ages about what *I* will do, what the experience and the ideas that *I have* that *I* will bring *to* the White House and *I* hope *to have* their *support* when *I*'m the Democratic nominee.' (H. Clinton, 17 January 2016)

> '*I think* now what *I've* called *for* is *counsel for* every *child* so that no *child has* to face any kind of process without someone who speaks and advocates *for* that *child* so that the right decision hopefully can be made.' (H. Clinton, 11 February 2016)

> 'Let's make the *Affordable Care* Act work *for* everybody.' (H. Clinton, 4 February 2016)

The specific sentences extracted from Sanders' answers explain clearly his positions with respect to Wall Street, some of the large companies, or on education. Moreover, the first and last examples tackle one of the recurrent topics for the Democrat, for which the terms 'college', 'university', and 'tuition' are specific in Sanders' rhetoric.

> '*Yes*, I do *believe* that now after the *American* people *bailed Wall Street* out, *yes*, they *should* pay a *Wall Street speculation* tax so that we can make *public colleges* and *universities tuition*-free.' (B. Sanders, 11 February 2016)

> 'Why does the *fossil fuel industry* pay, spend *huge* amounts *of* money on *campaign contributions*?' (B. Sanders, 11 February 2016)

> 'I do *believe* that *in* the year 2016 we have to look *terms of public education* as *colleges* as part *of public education* making *public colleges* and *universities tuition* free.' (B. Sanders, 11 February 2016)

One can complement this study by considering the terms ignored or used very infrequently in this primary election. For example, no selected candidate is discussing really the problem of the national debt. The word 'debt' appears in some utterances, but mainly in the context of the education debt for the Democrats. In the Republican party, the federal debt is debated by Rubio, and marginally by Kasich, and Paul.

When a question is discussed, the choice of the word can make the difference. With the immigration issue for example, Trump prefers using the term 'immigration' presenting this question more at an abstract level. On the other hand, Clinton could want to accentuate the human aspect and uses in this case the word 'immigrants'. Slight lexical differences can sometimes be important because, as mentioned by Lakoff and Wehling (2012), 'language is politics'.

6 Conclusion

This article has analyzed the style and the rhetoric used by the candidates during the 2016 US primary election. More precisely, this study has focused on the oral communication form using different TV debates in both political parties, a form less observed in previous studies.

During this primary election, Donald Trump presents clearly an atypical figure, employing short sentences, a reduced vocabulary, repeating the same arguments with simple words (see Table 2). When considering the most frequent lemmas, he is the single candidate to have the pronoun 'I' is the

second rank (after the article 'the'). The intensity of his ego can also be revealed by the fact that the most specific term in his dialogue is also the pronoun 'I' (see Table 6). In his answers, Trump prefers using intensively the verb construction (see Table 4), the pronoun 'I', and the negation (see Table 6). Among his most specific terms, one can see 'Mexico', and 'deal' reflecting two of his main concerns (immigration, and commercial trade agreements).

Hillary Clinton can also be characterized by a large use of the pronoun 'I' (fourth most frequent lemma, see Table 2) that is also over-used (see Table 6). None of the other Democrat candidates shows a clear intensive use of this pronoun. When considering overall stylistic indicators (see Table 3), Clinton, O'Malley, and Sanders present a high LD value as well as a higher number of BW and TTR ratio than the mean. Looking at her most specific sentences, Clinton tends to produce rather long sentences reflecting a more complex reasoning.

From overall stylistic measurements shown in Table 3, Ted Cruz appears with higher values than the mean. His answers contain more nouns and names adopting a more descriptive rhetoric. As depicted in Figs 1 and 2, Cruz represents a distinctive fraction of the Republican party. As one can see from his most specific terms and sentences, Cruz's concerns are related to a reform of the fiscal system, and the health care system.

Our findings must be confirmed by other studies comparing other electoral campaigns and taking into account the written form (e.g. electoral speeches, party manifestos, Web sites of the candidates, social networks information flow).

Acknowledgments

This research was supported, in part, by the NSF under Grant #200021_149665/1. The author wants to thank the anonymous reviewers for their helpful suggestions and remarks.

References

Argamon, S. and Levitan, S. 2005. Measuring the Usefulness of Function Words for Authorship Attribution. In *Proceedings of the 2005 ACH/ALLC Conference*, Victoria University Press, Victoria (BC), pp. 1–3.

Arnold, E. and Labbé, D. 2015. Vote for me. Don't vote for the other one. *Journal of World Languages*, **2**(1): 32–49.

Baayen, H. R. 2008. *Analysis Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Biber, G., Conrad, S., and Leech, G. 2002. *Longman Student Grammar of Spoken and Written English*. London: Longman.

Biber, C. and Conrad, S. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Bligh, M., Merolla, J., Schroedel, J. R., and Gonzalez, R. 2010. Finding her voice: Hillary Clinton Rhetoric in the 2008 presidential campaign. *Woman's Studies*, **39**(8): 823–50.

Boller, P. F. Jr. 2004. *Presidential Campaigns. From George Washington to George W. Bush*. Oxford: Oxford University Press.

Burrows, J. F. 2002. Delta: a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Cech, R., Garabik, R., and Altmann, G. 2015. Testing the thematic concentration of text. *Journal of Quantitative Linguistics*, **22**(3): 215–32.

Covington, M. A. and McFall, J. D. 2010. Cutting the Gordian Knot: the moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, **17**(2): 94–100.

Damerau, F. J. 1975. The use of function word frequencies as indicators of style. *Computers and the Humanities*, **9**(6): 271–80.

Francis, W. N. and Kucera, H. 1982. *Frequency Analysis of English Usage*. Boston, MA: Houghton Mifflin Co.

Greenacre, M. 2017. *Correspondence Analysis in Practice*. Boca Raton, FL: CRC Press.

Grieve, J. 2007. Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing*, **22**(3): 251–70.

Grimmer, J. and Stewart, B. M. 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, **21**(3): 267–97.

Hart, R. P. 1984. *Verbal Style and the Presidency. A Computer-based Analysis*. Orlando, FL: Academic Press.

Hart, R. P., Childers, J. P., and Lind, C. J. 2013. *Political Tone. How Leaders Talk and Why*. Chicago, IL: The University of Chicago Press.

Hewings, A., Painter, C., Polias, J., Dare, B., and Rhys, M. 2005. *Getting Started: Describing the Grammar of Speech and Writing*. Milton Keynes: Open University Press.

Kubát, M. and 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**(1): 33–80.

Labbé, C. and Labbé, D. 2006. A tool for literary studies. *Literary & Linguistic Computing*, **21**(2): 311–26.

Labbé, D. and Monière, D. 2003. *Le discours gouvernemental. Canada, Québec, France (1945-2000)*. Paris: Honoré Champion.

Labbé, D. and Monière, D. 2008a. *Les mots qui nous gouvernent. Le discours des premiers ministres québécois: 1960-2005*. Montréal, QC: Monière-Wollank.

Labbé, D. and Monière, D. 2008b. Je est-il un autre? In *Proceedings JADT 2008*, Presses universitaires de Lyon, Lyon, pp. 647–56.

Labbé, D. and Monière, D. 2013. *La campagne présidentielle de 2012. Votez pour moi!* Paris: L'Harmattan.

Lafon, P. 1980. Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, **1**(1): 127–65.

Lakoff, G. and Wehling, E. 2012. *The Little Blue Book: The Essential Guide to Thinking and Talking Democratic*. New York, NY: Free Press.

Laver, M., Benoit, K., and Garry, J. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review*, **97**(2): 311–31.

Lebart, L., Salem, A., and Berry, L. 1998. *Exploring Textual Data*. Dordrecht: Kluwer-Academic.

Lim, E. T. 2002. Five trends in presidential rhetoric: an analysis of rhetoric from George Washington to Bill Clinton. *Presidential Studies Quarterly*, **32**(2): 328–66.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, **19**(2): 313–30.

Millbank, D. 2016. Trump's Fake-News Presidency. *Washington Post*, November 18th.

Muller, C. 1992. *Principes et Méthodes de Statistique Lexicale*. Paris: Honoré Champion.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings 4th International AAAI Conference on Weblogs and Social Media*, AAAI Press, Palo Alto, CA, pp. 122–9.

Paradis, E. 2011. *Analysis of Phylogenetics and Evolution with R*. New York, NY: Springer.

Pauli, F. and 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. http://observer.com/2016/11/email-reveals-clinton-camp-spied-on-sanders-delegates-before-convention/.

Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology Evolution*, **4**(4): 406–25.

Savoy, J. 2013. The *Federalist Papers* revisited: a collaborative attribution scheme. In *Proceedings ASIST 2013*, ASIST, Montreal, QC.

Savoy, J. 2015a. Text clustering: an application with the *State of the Union* addresses. *Journal of the American Society for Information Science and Technology*, **66**(8): 1645–54.

Savoy, J. 2015b. Comparative evaluation of term selection functions for authorship attribution. *Digital Scholarship in the Humanities*, **30**(2): 246–61.

Savoy, J. 2015c. Vocabulary growth study: an example with the *State of the Union* addresses. *Journal of Quantitative Linguistics*, **22**(4): 289–310.

Savoy, J. 2015d. Authorship attribution using political speeches. In Tuzzi, A., Benesova, M., and Macutek, J. (eds), *Recent Contributions to Quantitative Linguistics*, vol. **70**. Berlin: De Gruyter, pp. 153–64.

Slatcher, R. B., Chung, C. K., Pennebaker, J. W., and Stone, L. D. 2007. Winning words: individual differences in linguistic style among U.S. Presidential and Vice Presidential candidates. *Journal of Research in Personality*, **41**(1): 63–75.

Stamatatos, E. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science & Technology*, **60**(3): 538–56.

Sylwester, K. and Purver, M. 2015. Twitter language use to reflects psychological differences between democrats and republicans. *PLoS One*, **10**(9): 1–18.

Tausczik, Y. R. and Pennebaker, J. W. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, **29**(1): 24–54.

Toutanova, K., Klein, D., Manning, C., and Singer, Y. 2003. Feature-rich Part-of-speech Tagging with a Cyclid Dependency Network. In *Proceedings of Human Language Technology - North American Chapter of the Association for Computational Linguistics*, ACL, Stroudsburg, PA, pp. 252–5.

Tuzzi, A. 2010. What to put in the bag? Comparing and contrasting procedures for text clustering. *Italian Journal of Applied Linguistics - Statistica Applicata*, **22**(1): 81–98.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. 2005. Opinion Finder: A System for Subjectivity Analysis. In *Proceedings Empirical Methods for Natural Language Processing*, ACL, Stroudsburg, PA, Vancouver, BC, pp. 34–35.

Young, L. and Soroka, S. 2012. Affective news: the automated coding of sentiment in political texts. *Political Communication*, **29**(2): 205–31.

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**(1): 33–48.

Yu, B. 2013. Language and gender in congressional speech. *Literary and Linguistic Computing*, **29**(1): 118–32.

Zhao, Y. and Zobel, J. 2007. Entropy-based authorship search in large document collection. In *Proceedings ECIR2007*. Heildelberg: Springer, LNCS #4425, pp. 381–39.