Glottometrics 38, 2017, 55-76

Analysis of the Style and the Rhetoric of the American Presidents Over Two Centuries

Jacques Savoy[1]

Abstract. To study the evolution of the rhetoric and language style of the American presidents from 1789 to 2017, we have analyzed all the annual *State of the Union* (SOTU) and inaugural addresses. Those speeches present the intentions and indicate the legislative priorities of the Chief of the Executive. Based on this relatively fixed form, this analysis corroborates several studies and emphasizes new trends and methods. With the years, the mean sentence length becomes shorter to facilitate comprehension by a larger audience. To further improve this goal, the vocabulary becomes less complex, and the percentage of big words decreases, indicating a preference for common words or expressions. The rhetoric evolves towards a more assertive and intimate tone with an increase in the occurrence of pronouns, particularly with the lemma *we*. Based on automatic classification techniques, we also observe that every president after 1961 generally has a clean and distinct style. When inspecting all presidents, some figures show a distinct break with their predecessors such as Lincoln, T. Roosevelt, Wilson, F.D. Roosevelt, or Kennedy. A closer analysis of presidencies since 1961 reveals that, with time, governmental speeches include more words related to humans and emotional terms, as well as references to God and symbolic expressions (*America*, *country*, *freedom*).

Keywords: *Governmental speeches, automatic classification, language models, authorship attribution, stylometry*.

1. Introduction

The presidential function has considerably changed from the time of the young republic under the presidency of Washington to Trump. To identify the broad trends of this evolution, several studies have analyzed presidential speeches. The number of such addresses remains low until the early twentieth century (Tulis, 1987), and their volume rose sharply after 1945 reaching an average of one speech per day under Carter's presidency (Hart, 1984). This evolution can be explained by the growing importance of journalists, media, and in particular, television. Beside their number, the content and the style of the presidential addresses has also evolved during the last two centuries. For political scientists J. Caesar et al. (1981), "speaking is governing" indicating the first role played nowadays by the governmental speeches. The president must convince the Congress as well as the citizens of his choice and persuade them (Neustadt, 1990) that the proposed policies are the most appropriate ones.

All addresses do not however have the same importance, and their type and audience vary. Faced with the impossibility of analyzing them all, past studies have limited their investigation to speeches considered as essential. For the United States, such analyses focus on the annual *State of the Union* (SOTU) allocutions and/or the inaugural addresses uttered during the swearing-in procedure. To supplement this selection, certain studies add some remarks delivered in a crisis context (e.g., address to the nation after the attacks of Sept. 11th, 2001).

[1] Computer Science Dept., University of Neuchatel, Switzerland, Jacques.Savoy@unine.ch

Based on both past studies and using our proposed methods, this article describes the main changes in the US presidential rhetoric and style from Washington to Trump using computational tools. In this view, rhetoric is defined as the art of effective and persuasive speaking, the way to motivate an audience, while language style is present as pervasive and frequent forms used by an author for mainly aesthetical value (Biber & Conrad, 2009).

The rest of this paper is composed as follows. The second section outlines the value and characteristics of the SOTU and inaugural speeches. To analyze this corpus, the third section describes previous studies and the main methods used. In the fourth, a set of overall stylistic measurements describes the main trends of the presidential rhetoric evolution. In the fifth section, the part-of-speech (POS) distribution is used to differentiate the presidents' styles. In the sixth section, automatic classification is applied to generate a map showing the differences between presidents based on their stylistic preferences. The last section provides a more detailed analysis of presidential rhetoric and style since Kennedy (1961).

2. Selection of the Governmental Speeches

To analyze the presidential rhetoric and its evolution, the majority of studies are based on a subset of the *State of the Union* (SOTU) addresses. Using computer technology, the entire set can be analyzed which includes 228 speeches given by 43 presidents from Washington (Jan. 8th, 1790) to Trump (Feb. 28th, 2017). This SOTU address is required by the US Constitution (Article II, Section 3) where it is mentioned that the president must provide information to the Congress about the state of the Union and "*measures as he shall judge necessary and expedient*". Such an address provides both an analysis of the current situation, indicates the president's priorities and presents the legislative agenda for the coming year. All of them are available on the internet (e.g., www.presidency.ucsb.edu or MillerCenter.org) and an annotated version of the 20th century addresses was recently published (Kalb & Peters, 2007).

The following reasons explain the importance of the *State of the Union* (SOTU) addresses. First, the United States occupies a position of global importance, consequently the president's vision transcends the interests of a single country. Second, this set of governmental allocutions covers a period spanning more than two centuries allowing us to analyze the evolution of the rhetoric and style. Moreover, they are delivered in a relatively stable institutional context, reducing some factors of variation. Fourth, several of these speeches have outlined significant political positions such as the Monroe Doctrine (1823), the four freedoms (F. D. Roosevelt in 1941), or the war against poverty (L. B. Johnson in 1964). More recently, this allocution was an opportunity to introduce new phrases such as "*axis of evil*" (G. W. Bush in 2002). Several studies describe in detail the institutional and political context of these speeches (Kolakowski & Neale, 2006), (Hoffman & Howard, 2007), (Shogan & Neale, 2012).

As a second major source of US government speeches, previous studies have considered the 58 inaugural addresses uttered at the beginning of each term by each of the 40 elected presidents. This set begins with the first allocution (Apr. 30th, 1789) uttered by Washington and ends with the allocution of Trump (Jan. 20th, 2017). For all of them, the oral form was chosen and the general form and topics are more diverse than with the SOTU speeches. They often present the main objectives for their term in the White House, expose their intentions regarding foreign policy, and give broad guidelines fixed for the new administration. However, their lengths show some variability. For example, the second inaugural speech of Washington (Mar. 4th, 1793) includes only

four sentences (145 words) while that of W. Harrison (Mar. 4th, 1841) was the longest with 8,356 tokens. A complete list of the selected speeches is provided in Table A.1 in the Appendix.

One can consider that (oral) speeches delivered by the presidents correspond to an oral communication form while (written) messages (e.g., sent to the Congress) must be categorized as a distinct text genre. However, as mentioned by Biber & 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".

If one considers the president as the author of a speech, one does not take this literally. It is known that behind each important politician one can usually find a team of ghostwriters (Humes, 1997). For example, under Kennedy's presidency, the main ghostwriter was Sorensen (Carpenter & Seltzer, 1970), Madison & Hamilton behind Washington, and Favreau[2] & Keenan behind Obama. However, though some presidents were actually the author of their speeches (e.g., Lincoln), the tenant of the White House is involved more or less intensively in drafting their important speeches (Hume, 1997). The website *YouTube.com*[3] provides some video showing the preparation of some of Obama's SOTU addresses.

Finally, this analysis ignores two tenants of the White House (W. Harrison (1841), J. Garfield (1881)) because they just uttered one inaugural allocution, without any SOTU addresses, as their terms were limited to a few months. Moreover, in this study one can find only two speeches uttered by Trump (with a length of 6,252 tokens, the smallest over all remaining presidents); therefore, the findings related to the last US president must be taken with prudence.

3. Related Work and Methods

To analyze the rhetoric and style of presidential writings, the first quantitative linguistics studies focused on the word usages and their frequencies (Baayen, 2008, Jockers, 2014, Popescu et al., 2009). 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 leads to similar conclusions. If the definition of a lemma is clear, the term "word" is usually ambiguous. The expression word token (or simply token) refers to an occurrence or instance of a word type (or type). For example, the sentence "I saw a man with a saw" counts seven tokens for five word types (I, saw, a, man, with), and six lemmas (I, (to) see, a, man, with, saw). In this study, the statistics are usually computed based on lemmas.

For Biber & Conrad (2009), a stylistic study should be based on ubiquitous and frequent forms. As an operational definition, our analysis is based on the k most frequent word types or lemmas, with k = 200 or 300, values that have been justified in author attribution studies (Burrows, 2002), (Savoy, 2015a).

Simple frequency analysis may report interesting aspects of the evolution of presidential style. For example, Figure 1 illustrates the evolution of the relative fre-

[2] J. Favreau comments his ghostwriter job at www.youtube.com/watch?v=zFbaesLEa4g

[3] See at www.youtube.com/watch?v=FxwcJx0-21E the behind the scene of *State of the Union* address of 2012, or at www.youtube.com/watch?v=BaR2jXboVQ0 for the SOTU 2014.

quency of the definite determiner *the* and the lemma *we*. Until Taft's presidency, some stability can be observed with a maximum for the determiner reached by Q. Adams (9.9%). For the pronoun *we*, a maximum is reached under Carter's presidency (4.7%), and this frequency remains relatively high and stable after this extreme value.

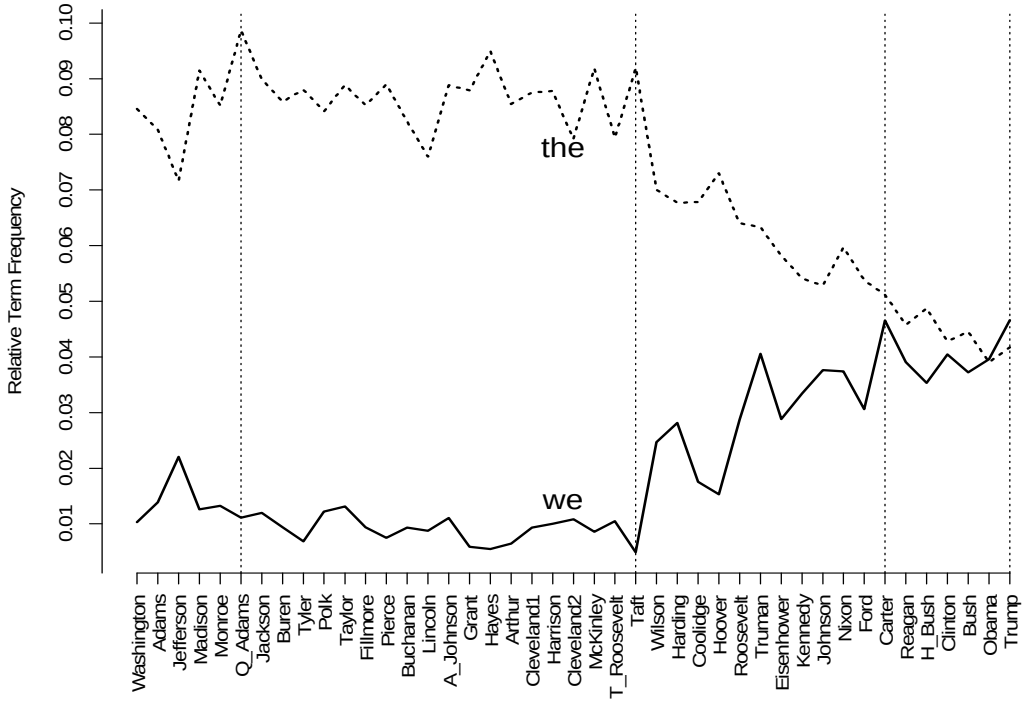

Figure 1. Evolution of the relative occurrence frequency of the lemmas *the* and *we*

In addition, a simple study may focus on the mean sentence length (number of tokens). The presence of long sentences indicates a substantiated reasoning or specifies the presence of detailed explanations. Even if a long sentence is required, its length is usually not conducive to an easy understanding.

Word length is another indicator of a message's complexity (Hart, 1984), (Tausczik & Pennebaker, 2010), the longer the words, the higher the complexity. Of course, a simple count of the number of letters to indicate the word complexity should be taken with caution. The letters are not the direct constituent of the word (a role played by the syllables or the morphemes). Moreover, the graphophonic relationship is not direct and simple in the English language. Nevertheless, Lakoff and Wehling's study (2012) indicates a relationship between the word length and word complexity analyzed be y the receiver.

> "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 & Wehling, 2012, p. 41).

Based on this finding, the frequency of word types composed of six letters or more indicates the use of a rich and sophisticated vocabulary (Tausczik & Pennebaker, 2010). Of course, this limit of six letters is arbitrary and another value can be used with

an alphabet-based language. For the Chinese language based on sonograms, or for the Korean based on syllabic system (called Hangul), this limit can be fixed at one or two characters[4]. In fact, Lee et al. (1999) found more than 80% of Korean nouns were composed of one or two Hangul characters, and for Chinese, Sproat (1992) reported a similar finding.

In conclusion, a text or a dialogue with a high percentage of big words tends to be more complex to understand. L. B. Johnson recognized this rhetoric problem by specifying to his ghostwriters: "I want four-letter words, and I want four sentences to the paragraph." (Sherrill, 1967).

Our two last measurements (i.e. mean sentence length and word length) are not fully independent and a relationship does exist between them, known as the Arens' law (Grzybek et al., 2008). However, the words are not the direct constituents of a sentence but only through phrases or clauses. Therefore, the relationship is less direct than expected.

To complement these frequency measures, Muller (1992) demonstrates the significance of studying the specific vocabulary for a given author (or a period), by defining a measure of lexical specificity. This approach was used for analyzing the political vocabulary in France, Quebec, and Canada (Labbé & Monière, 2003; 2008). Using this measure, the lexical specificity of the different French presidents or prime ministers can be detected. Moreover, this measure is not limited to words, but can include punctuation marks, or larger components such as word bigrams, trigrams or noun phrases (Savoy, 2010; 2016).

Other studies focus more on the Part-Of-Speech (POS) categories and their distribution. In this case, the automatic morphological tagging software (Manning & Schütze 1999) can be used to assign to each word token a syntactic and morphological label. For example, in the tagged sentence "It/PRP begins/VBZ with/IN our/ PRP$ energy/NN" we find labels attached to nouns (NN, common noun, singular, NNS common noun, plural), proper names (NNP), verbs (VB dictionary entry, VBZ for present tense, 3rd person, etc.), adjectives (JJ), pronouns (PRP), prepositions (IN), or adverbs (RB). This information can be employed to impose useful constraints for generating noun-phrases (bigrams or trigrams). For example, the pattern JJ-NN or NN-NN can extract the bigrams *clean energy*, *new jobs* or *nuclear weapons*. Considering only the most frequent bigrams results in uninteresting patterns such as *of the*, *in the*, …

Finally, several studies have suggested measuring the distance between two texts based on a predefined list of words, or the *k* most frequent word types or lemmas (with $k = 50$ to 1,000, (Burrows, 2002)) or simply according to the whole vocabulary (Labbé, 2007). Such measures have been used to determine the real author of a given document or text excerpt (Labbé, 2007), (Savoy, 2014). Knowing the distance between documents, an automatic classification algorithm (Baayen, 2008), (Jockers, 2014), (Arnold & Tilton, 2015) can visualize similarities between texts (or sets of texts) written by the same author. This allows one to draw maps showing the stylistic similarities between presidents (as shown in Section 6) or draw graphic affinities according to the main topics of their speeches (Savoy, 2015b). Other studies offer similar approaches to detect and monitor topics over time (Rule et al., 2015) or across scientific journals (Mimno, 2012).

[4] Instead of considering directly the sinogram or the Hangul character, one can count the number of strokes required to draw the corresponding character.

As another form of rhetoric analysis, several studies attempt to regroup several word types under a semantic tag. For example, under the category *Symbolism*[5], the DICTION system (Hart, 1984), (Hart et al., 2013) includes a list of words related to country (e.g., *America, nation*), ideology (e.g., *freedom, peace, rights*), or generally, political institutions or concepts (e.g., *government, law*). In a similar way, the system LIWC (Linguistic Inquiry & Word Count) (Tausczik & Pennebaker, 2010) regroups terms under syntactical, emotional or psychological categories. Such word classes may correspond to specific grammatical categories (e.g., first person singular denoted *Self* (*I, me, mine, my*)), broader ones (e.g., personal pronouns) as well as more complex ones (verbs in the future tense, auxiliary verbs). With some semantics, the LIWC system defines the inclusive class (under the label *Incl*) (e.g., *add, and, both*) or exclusive category (*Excl*) (e.g., *except, or, but, without*). As more pertinent categories, we may encounter positive emotions (*Posemo*) (e.g., *happy, hope, peace*) or negative (*Negemo*) (e.g., *humiliat*, war*), *Cognition* (e.g., *admitted, perceive*) or terms related to *Human* (e.g., *family, friend, child**). More details are given in (Tausczik & Pennebaker, 2010), (Hart, 1984). Those semantic categories will be used in Section 7 to analyze the rhetoric of the presidencies since 1961.

As examples of the usefulness of such categories, we reproduce below a passage having a high density in the category *Posemo* (words in italics, G. W. Bush), and a second text excerpt showing a high degree in the category *Affect* (Reagan).

> "For the *brave* Americans who bear the risk, no victory is *free* from sorrow. This Nation fights reluctantly, because we know the cost and we dread the days of mourning that always come. We seek *peace*. We strive for *peace*. And sometimes *peace* must be defended." G. W. Bush, *State of the Union*, Jan. 28th, 2003.

> "People of the Soviet, President Dwight Eisenhower, who *fought* by your side in World *War* II, said the essential *struggle* "is not merely man against man or nation against nation. It is man against *war*." Americans are people of *peace*. If your government wants *peace*, there will be *peace*. We can come together in *faith* and *friendship* to build a *safer* and far *better* world for our children and our children's children." R. Reagan, *State of the Union*, Jan. 25th, 1984.

4. Overall 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 & Manning, 2000) 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 43.9 tokens/sentence while the mean over all presidents is 33.3, as depicted in the last column of Table A.2 (shown in the Appendix). With Trump, the mean sentence length decreases to 20.5 tokens/sentence. These examples, together with Figure 3 (see below), clearly indicate that the style is changing over time. Currently, the preference goes to a shorter formulation, simpler to understand for the audience.

[5] In this study, measures, rhetorical, or stylistic indicators are shown in italic and are capitalized.

As a second global stylistic measurement, the frequency of big words (composed of six letters or more, and denoted BW) can be analyzed (Tausczik & Pennebaker, 2010). A text or a dialogue with a high percentage of big words tends to be more complex to understand. With this measurement, Eisenhower has the highest value (36.1%) over all presidents while G. H. Bush presents the lowest (25.6%) (values depicted in Table A.2 in the Appendix). As a general trend, one can see that recent presidencies (from Reagan) tend to employ less big words, in an attempt to simplify their formulations and produce less complex explanations.

Figure 2 depicts these first two stylistic measurements with the mean sentence length (y-axis, computed according to the number of tokens) and the percentage of big words in the addresses delivered by all presidents (x-axis). On the top, one can find the Founding Fathers (with Madison depicting the highest mean sentence length (48.3 tokens/sentence)), and below them the presidents of the 19th century. On the bottom left, we see the contemporary presidents (Obama, G. H. Bush, and Clinton) opting for short sentences and few big words. On the extreme right with a large percentage of big words, we discover Hoover (1929-1933) and Eisenhower (1953-1961) depicting the largest percentage of big words (36.1%). Trump presents the second smallest mean sentence length (20.5 tokens/sentence), slightly more than G. H Bush (18.9).

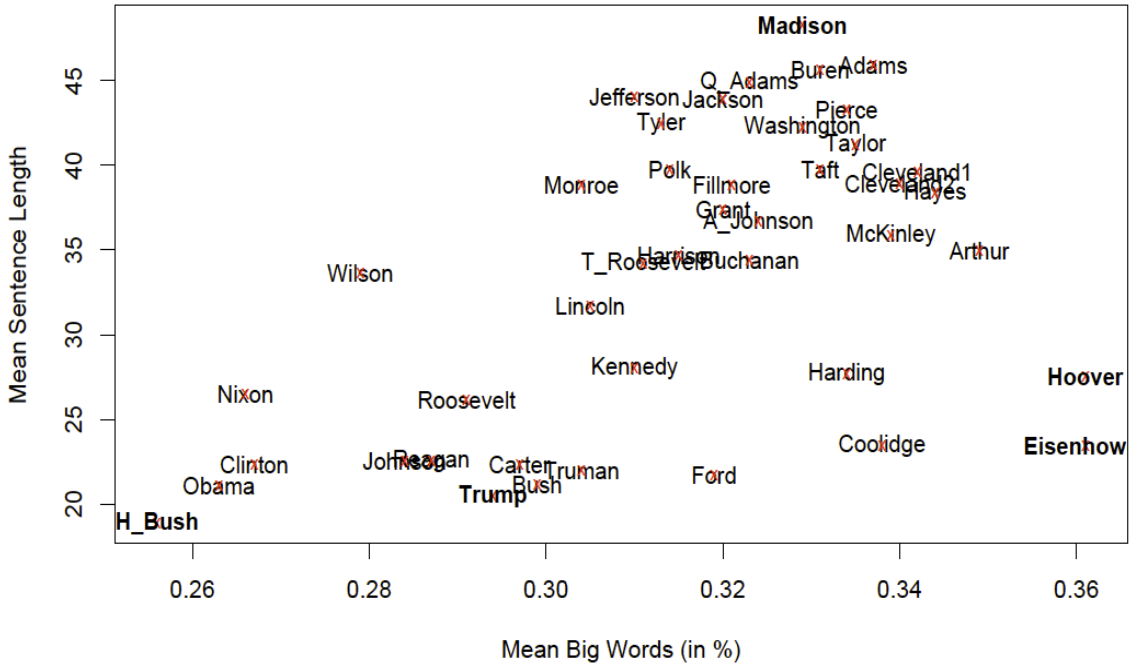

Figure 2. Relationship between mean sentence length (MSL, y-axis) vs. percentage of big words (BW) (> 5 letters, x-axis) based on the *State of the Union* and inaugural addresses

As a third stylistic indicator, the lexical density (denoted LD) 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 273 entries. As depicted in Equation 1, this LD value is given in percentage over the number of tokens.

$$LD(t) = \frac{lexical\ word(t)}{n(t)} = 1 - \frac{functional\ word(t)}{n(t)} \quad (1)$$

A relatively high LD percentage indicates a more complex text, containing more information. Over all presidencies, the LD values varies from 50.3% (Eisenhower) who focusses more on topical forms and expressions to a minimal value of 41.5% (Wilson). For the last presidents (since 1980), this value is relatively stable and around 47% as shown in Figure 3.

The TTR (Type-Token Ratio) or the relationship between the vocabulary size and the number of word types (Baayen, 2008), (Popescu et al., 2009), (Mitchell, 2015) 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 unstable, tending to decrease with text length (Baayen, 2008). To avoid this problem, a better computation is provided in (Covington & McFall, 2010) or (Popescu, 2009) that suggest taking the moving average of TTR denoted MATTR. This computation technique has been adopted.

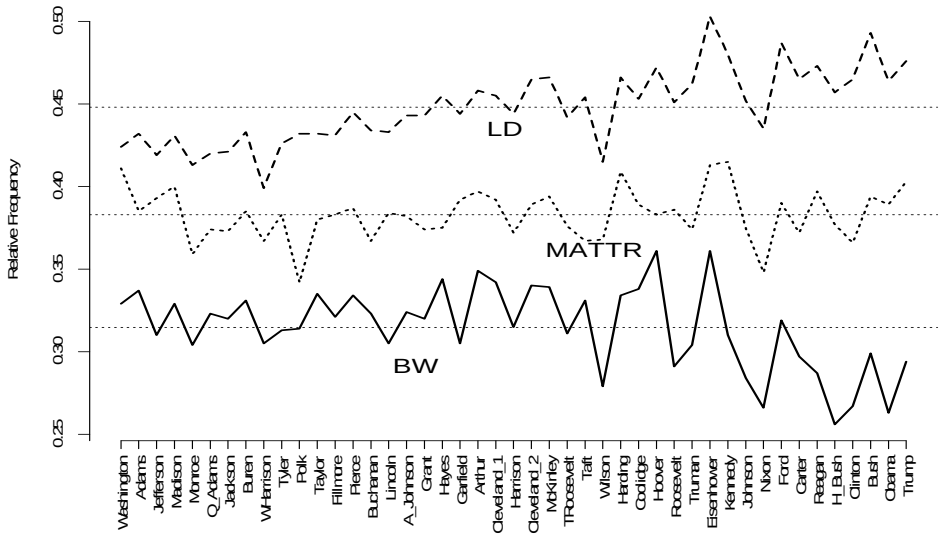

Figure 3. Evolution of lexical density (LD), big words (BW) and moving average type-token ratio (MATTR)

An overview of the values of these four stylistic indicators is reported in Table A.2 for some presidents. In this table, the largest values are depicted in bold, while the smallest are shown in italic. From this data, Washington exhibits a high MATTR value (41.2%) compared to Wilson (1913-1921) (36.8%). As other presidents presenting a rich vocabulary, one can mention Kennedy (41.2%) or Eisenhower (41.1%). On the opposite side, Wilson employs a simple vocabulary with a low MATTR value (36.1%), as well

as Nixon (34.8%) or Monroe (35.9%). For this stylistic measurement, the evolution is not related to the time but corresponds more to a personal choice.

5. Part-Of-Speech Distribution

After applying the Stanford POS tagger (Toutanova et al., 2003) over the inaugural and SOTU addresses, Table 1 shows the percentage of each grammatical category for some selected presidents. In this table, the largest values are depicted in bold, while the smallest are shown in italic. The last line under the label "Other" regroups other categories such as punctuation marks, numbers, symbols, dollar signs, or foreign words.

A first analysis indicates that nouns and adjectives represent a high percentage of Eisenhower's addresses. This emphasis on nouns can be justified by a real need for explanations. A lower frequency characterizes Lincoln's or Trump's speeches. Names (second row) are used frequently by Trump. Pronouns are very frequently used by Clinton and Obama, especially with the lemma *we* (see Figure 1). The verb and adverb categories also occur frequently in Obama's speeches, indicating that these remarks are more oriented towards action (these two findings give us a new perspective on the sentence: "*yes, we can*"). To a lesser extent, this finding can be applied to Clinton while Kennedy (JFK) tends to use this part-of-speech less. A high usage of verbs indicates dynamic thinking characterizing a person who analyzes a new problem from its historical forces or developing perspective (Pennebaker, 2009). A high occurrence frequency of determiners and prepositions characterizes the first presidency (Washington), a finding that can also be seen in Figure 1. These two categories characterize longer syntactic constructions. Since the '80s (Reagan's presidency), their use decreases and the adopted style favors shorter sentences with a higher occurrence of pronouns. This facet corresponds to a more direct tone, trying to establish a close personal relationship with the audience.

Table 1
Percentage of various POS for some selected presidencies

	Wash.	**Linc.**	**Wilson**	**Roos.**	**Eisen.**	**JFK**	**Reagan**	**Clinton**	**Obama**	**Trump**
noun	19.9	*18.2*	19.8	20.9	**22.7**	21.5	20.1	19.5	19.6	18.7
name	3.0	4.1	*1.8*	3.0	3.6	3.4	3.9	3.9	3.5	**5.8**
pron.	5.7	*4.8*	7.8	6.7	5.5	6.5	8.2	**9.5**	9.1	8.9
adj.	7.4	7.8	7.9	8.5	**9.4**	8.4	7.5	7.0	*6.5*	6.6
verb	14.9	14.6	15.0	13.8	13.3	*12.8*	14.9	15.4	**16.5**	15.1
adverb	3.8	5.0	4.7	4.1	*3.8*	4.2	4.8	4.6	**5.2**	4.7
det.	**12.9**	12.3	11.0	11.0	10.3	10.1	9.0	8.9	8.8	*8.1*
prep.	**19.3**	17.6	17.5	16.7	15.7	14.7	14.0	14.0	13.3	*12.5*
coor.	*3.5*	4.0	**4.9**	4.1	3.7	4.7	4.1	3.8	4.1	4.5
other	9.5	11.5	9.5	11.3	12.0	13.7	13.6	13.5	13.3	15.1

To verify how each presidency can deviate from an expected average style, we generate a centroid distribution over all POS tags by computing the mean distribution over all 43 presidents. Using the chi-square test (Conover, 1971), we found that each presidency deviates significantly ($p < 0.001$) from this centroid distribution. The closest presidency to this mean profile is J. Adams (1797-1801).

To better visualize the relationships between the POS categories and the presidents, a principal component analysis (PCA) (Baayen, 2008) has been applied on the data depicted in Table 1 (with some additional presidencies) using the R software (Jockers, 2014). In Figure 4, the horizontal axis emphasizes the contrast between the pronouns (and punctuation) appearing on the extreme right, and a group composed by a combination of prepositions and determiners (the two labels are also superposed) depicted on the left. At the ends of this axis, we can observe the opposition between Trump (and, to a lesser extent, Clinton and G. H. Bush) on the one hand and, on the other hand, T. Roosevelt (and, in part, Washington). As indicated, this first principal component axis represents 44.1% of all variability, while the vertical axis adds 19.3%. Thus, Figure 4 illustrates 63.4% of the total variance. Along this second axis, the verb appears on the top with Jackson and Obama as representatives. On the bottom, we encounter the group of nouns and adjectives, with Kennedy and Eisenhower as figureheads.

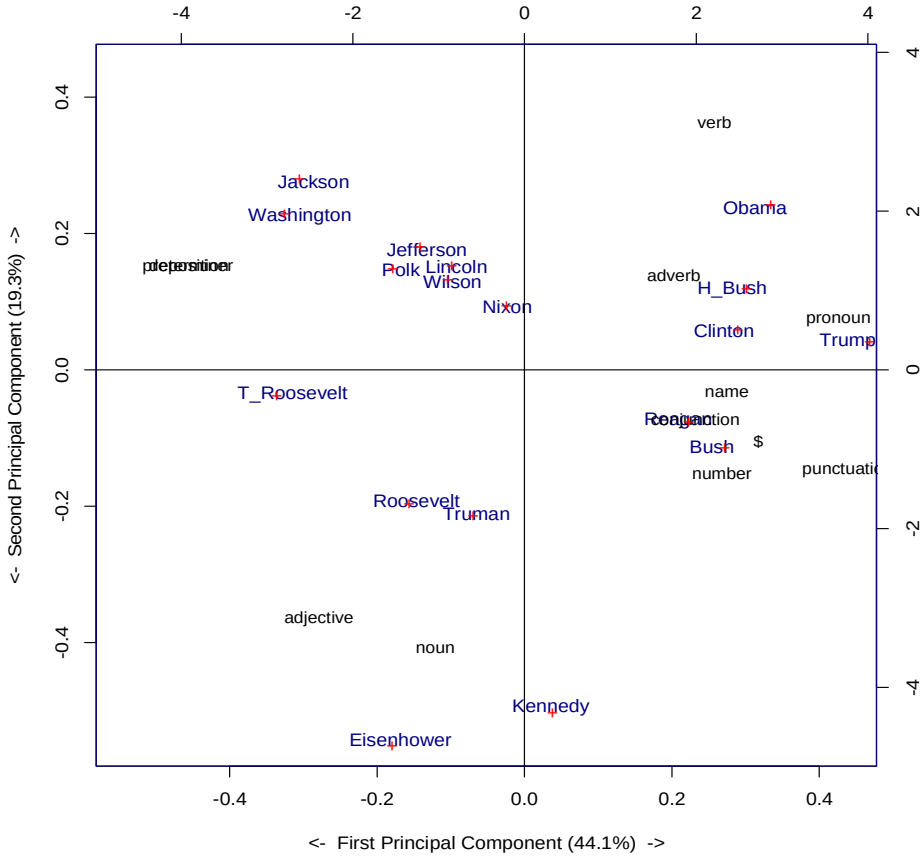

Figure 4. Principal component analysis (PCA) based on the POS distribution based on the *State of the Union* and inaugural addresses.

On the right part of Figure 4, we can observe the recent presidents with Clinton, G. H. Bush, Obama, and Trump, and just below Reagan, and G. W. Bush. These presidencies are using pronouns, verbs, adverbs, numbers, and punctuation symbols more frequently (their sentences are shorter, implying more full stops).

On the top left part, the early presidencies are regrouped with Washington, Jefferson, and Jackson, using more determiners and prepositions. The average speech, located in the center of the figure, does not have a clear representative. In this neighbor-

hood, one can see Nixon, and Wilson. Figure 4 also indicate that Eisenhower's and Kennedy's presidencies were clearly different from the others. Both are located on the bottom corresponding to speeches presenting more noun-phrases (nouns and adjectives).

Finally in this figure, the affinities between presidents do not follow a political party affiliation. For example, we find groups formed by a representative of each party (e.g., G. H. Bush – Clinton or Eisenhower – Kennedy). The subdivision into clusters seems to correspond better to a temporal proximity, a finding confirmed by Rule et al.'s study (2015).

6. Automatic Clustering Based on Stylistic Features

To globally compare the different presidential styles, the intertextual distance between two texts have been computed according to Labbé's method (2007). With this metric, the returned value depends on the overlapping between the two texts and varies between 0.0 and 1.0. Between these two extremes, the distance depends on the number of lemmas in common on both texts and their frequencies. In this study to represent each presidency, a profile is generated by concatenating all addresses delivered by a given president.

According to Labbé's definition, the intertextual distance between Profile A and Profile B is given by Equation 2 in which V_A (or V_B) indicates the vocabulary of Profile A, tf_{iA} (respectively tf_{iB}) denotes the term occurrence frequency of the ith word type in Profile A, and n_A (respectively n_B) the length of Profile A (number of tokens).

$$dist(A,B) = \frac{1}{2n_A} \sum_{i \varepsilon V_A \cup V_B} |tf_{iA} - tf_{iB}| \quad (2)$$

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 Profile B) to the size of the smallest one (Profile A in our example). To achieve this, the term frequency of each word type belonging to the largest text is modified as follows:

$$tf'_{iB} = tf_{iB} \cdot {}^{n_A}/_{n_B} \quad (3)$$

To reflect only the stylistic aspects, each profile is represented by the top 300 most frequent lemmas occurring in the SOTU and inaugural addresses. Applying this distance measurement for each pair of profiles, we obtain a symmetric matrix (43 x 43 = 1,849 values). Just showing all these values does not provide a useful picture. On the other hand, this information can be used to apply an automatic classification (Baayen, 2008). The result, depicted in Figure 5, allows us to discover the clusters generated based on similar stylistic profiles.

In Figure 5, the distance between each president is visualized by a technique derived from genomic trees (Paradis, 2011), more precisely using the nj() function (Rzhrtsky and Nei, 1993), (Gascuel and Steel, 2006) available in R (Jockers, 2014). In this picture, the line length joining two presidents is proportional to the distance between them. For example, to go from Lincoln to Kennedy, we must travel a greater distance than between Lincoln and Roosevelt. Finally, to be on the left or the right, up or down, does not matter. This position is selected to allow a better overall visualization.

In this figure, the longest distance (0.337) can be found between Obama and Q. Adams, and the second largest (0.334) links Clinton to Q. Adams. The shortest distance (0.066) connects Jackson with his successor van Buren, while the second (0.075) joins the two Cleveland presidencies (1885-1889 and 1893-1897).

Following a movement from bottom to top, the presidents are placed almost in chronological order. On the bottom of the figure, the first group includes the Founding Fathers (Washington, Adams, Jefferson, Madison, and Monroe). This group covers the period from 1790 to 1825. A closer inspection reveals that Monroe is a little bit further apart from the kernel formed by the first four presidents. Although the stylistic distance is still small among these five presidents, the political vision of Jefferson or Madison (limited federal power) is different from that shared by Washington and J. Adams (strong federal government).

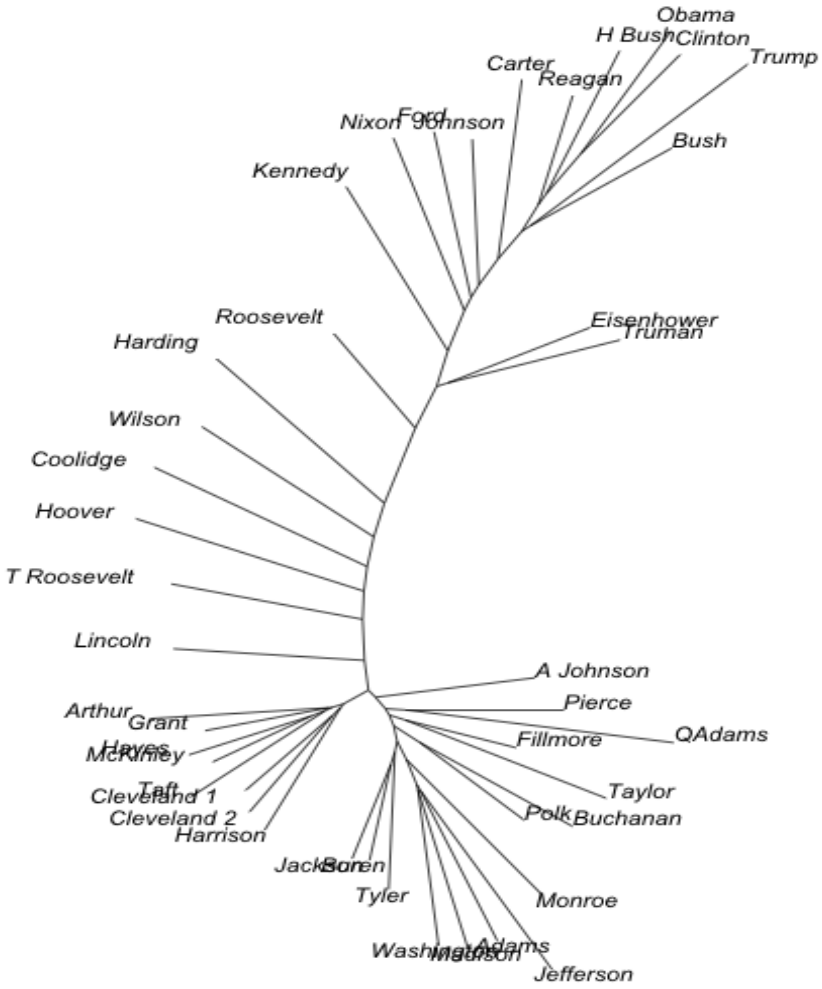

Figure 5. Tree representation of stylistic distances between the president profiles (with k = 300 most frequent lemmas) based on the *State of the Union* and inaugural addresses.

The second period comprising the years 1829-1845 is located just on the left, formed by the Democratic trio Jackson-Van Buren-Tyler. On the right, and closely related, we can find the cluster Polk (1845-49) - Buchanan (1857-1861) and the Whig duo Taylor-Fillmore (1849-1853). Just above, the pair Q. Adams (1825-1829) and Pierce (1853-1857) share a common style but with some temporal distance. Finally, one can see A. Johnson (1865-1869) as an isolated figure having a district style.

On the left, beginning with Arthur (1881-1885), we can find a group of presidents (Grant, Hayes, McKinley, Taft, Cleveland, and B. Harrison) covering the period 1869-1913. Except for the two terms of Cleveland (1885-1889, 1893-1897), all are Republicans and form a fairly homogeneous group based on stylistic considerations.

Above this large cluster, we find a sequence of presidents having a style significantly different from each other. The first in this series is Lincoln (1861-1865) who owns a style usually considered as the most beautiful. In Figure 5, this president is clearly located further away from presidents covering the same time period. A. Johnson (1865-1869), who succeeded Lincoln, also presents a particular style that is closer to his predecessors. From a stylistic evolution point of view, A. Johnson's presidency marks a step backward.

Within the first half of the 20th century, three presidents clearly stand out from a stylistic point of view. First, T. Roosevelt (1901-1909) depicts a large distance with his predecessor (McKinley) and his direct successor (Taft), both appearing in the group of presidents covering the period 1869-1913. Second, Wilson (1913-1921) modernizes the presidency; the United States become a great power wishing to play a global role (Nye, 2013). Wilson adopts a clearly distinctive style to help him in achieving this goal. The third innovative president is F.D. Roosevelt (1933-1945) whose style stands out clearly from its predecessors (Coolidge and Hoover) both located between T. Roosevelt and Wilson. Harding (1921-1923) also appears as owning a distinct style compared to the others, but his presidency was judged as one of the worst in US history (Riding et al., 1997).

From 1946, each presidency depicts a pretty distinct style that the next section will describe with more detail. As exceptions, we can find the binomial Truman-Eisenhower or the strong similarity between Clinton and Obama. A quick inspection reveals that the presidents appear in almost perfect chronological order; the first exception is L. B. Johnson appearing after Nixon-Ford, the second is G. W. Bush located closer to Carter, and the third with Trump who is located close to G.W. Bush. Figure 5 also highlights the impact of the style of Kennedy, inaugurating a brighter presidential style located at a farther distance from his predecessors.

7. Analysis of the Presidential Styles Since 1961

This last section analyzes with more detail the rhetoric and presidential style since 1961 (Caesar et al., 1981; Hart, 1984; Tulis, 1987; Neustadt, 1990; Gelderman, 1997; Kalb & Peters, 2007; Greenstein, 2009). To limit this analysis, Ford and Carter's presidencies will be ignored. The main indicators used in our comparison are reported in Table 2, with the mean over these nine presidencies indicated in the last column.

In this analysis, a selected set of semantic categories defined by the DICTION (Hart, 1984; Hart et al., 2013) or LIWC system (Tausczik & Pennebaker, 2010) have been used. Such lists may regroup specific grammatical categories such as *Self* defined by the word tokens (*I, me, my, mine*), tokens related to a given topic (e.g., *Human* with (e.g., *child**, *family*, *friend*), or *Social* with (e.g., *societ**, *speak*, *tell*, *team*)), terms denoting an emotion (*Posemo* with (e.g., *hope*, *win*, *best*), *Negemo* with (e.g., *fear*, *tear*, *sadness*)) or other rhetoric aspects such as *Cognitive* mechanism (e.g., *cause*, *think*, *organize*, *realis**), *Concreteness* words (e.g., *bank*, *college*, *troop*, *police**) or *Tentative* language forms (e.g., *maybe*, *perhaps*, *appear*).

From these indicators, one can build a centroid reflecting the average president by computing the mean over all measurements. Applying the chi-square test (Conover, 1971), we found that each presidency described in Table 2 deviates significantly

($p < 0.001$) from the centroid distribution. The closest to this average presidency is Reagan.

Kennedy's presidency corresponds to an intellectually brilliant, but impersonal, style. His rhetoric will inspire and motivate a nation and, thanks to the absence of excessive patriotism (the lowest *Symbolim* value: 1.8%, see Table 2), this motivation wins throughout the Western world. For Pennebaker (2011), Kennedy owns a complex thinking, able to convey complex problems and ideas in a rhetoric that the people can understand (highest value in the category *Tentative*: 1.8%, see Table 2). As reported in Table 1, this presidency presents a higher percentage of noun phrases (nouns and adjectives), and a low percentage of verb phrases (verbs and adverbs). Kennedy's tone is also reflected by the frequent use of words belonging to the *Exclusive* category (e.g., *but, rather, without*) (2.4% indicated in Table 2, with a mean over the nine presidencies of 2.1%), *Causal* (e.g., *because, effect*), and negation (e.g., *not, never*, 1.4%, mean: 1.2%). On simple measures, the Kennedy presidency is also characterized by greater complexity (longer sentences and a relatively high percentage of big words as shown in Figure 3). His Type-Token Ratio (MATTR) is higher (41.5%) than the average (38.3% reported in Table A.2) indicating a richer vocabulary. As for Truman, the lemma *we* (*we, us, ours*) is significantly over-used (see Figure 1) compared to previous presidents.

Table 2
Percentages of different semantic categories for some presidents

	JFK	Johnson	Nixon	Reagan	H Bush	Clinton	Bush	Obama	Trump	Mean
Self	*0.7*	1.6	1.3	1.1	**2.0**	1.5	0.9	1.2	0.9	1.2
Incl.	9.0	9.5	*8.6*	9.5	9.7	9.5	10.3	9.3	**11.0**	9.6
Excl.	2.4	1.9	2.0	2.0	2.2	2.1	*1.5*	**2.6**	1.8	2.1
Negat.	1.4	1.0	1.1	1.3	1.2	*1.0*	*1.0*	**1.5**	1.1	1.2
Symbo	*1.8*	2.4	3.2	2.8	2.4	2.5	3.1	2.2	**3.4**	2.6
Cogn.	19.5	20.1	*18.3*	19.4	**21.0**	20.4	20.2	20.7	20.5	20.0
Humar	*5.7*	7.1	6.4	7.5	7.5	8.8	7.9	7.9	**9.2**	7.6
Concr.	4.3	5.0	*4.3*	4.7	4.5	5.3	5.1	5.4	**6.1**	5.0
Tenta.	**1.8**	1.7	1.4	1.4	1.5	1.5	1.4	**1.8**	*0.8*	1.5
Social	*8.4*	9.8	9.3	10.4	10.8	12.1	11.4	10.8	**12.4**	10.6
Posemo	4.2	*3.9*	4.4	4.9	4.4	4.4	**5.2**	4.0	4.7	4.5
Negem	2.3	1.9	*1.6*	1.8	1.7	*1.6*	**3.0**	1.7	2.0	1.9

With L. B. Johnson, the presidential style becomes more direct, simple, and popular. His lexicon complexity goes down to a LD value of 45.2% (compared to 48% for Kennedy or 50.3% for Eisenhower). At the level of personal pronouns, we can detect a clear increase in the first singular pronoun (*I, me, mine*, 1.6% as reported in Table 2), a characteristic that will also be overserved under G. H. Bush's presidency.

With Nixon, the presidency becomes imperial according to the bestselling title "*The Imperial Presidency*" (written by A. M. Schlesinger). The rhetoric becomes clearly assertive, optimistic (*Posemo* of 4.4% comparted to 3.9% for Johnson), and relies on familiar words. The MATTR value is 34.8%, the second lowest value over all US presidents (Polk presents the minimal MATTR value with 34.2%). Moreover, the LD value decreases to 43.5% compared to 45.2% for Johnson. Another facet of Nixon's style is the recurrent use of terms belonging to the *Self* category (shown in italics in the following example):

> "*I* know these have no ideology, no race. *I* know America. *I* know the heart of America is good. *I* speak from *my* own heart, and the heart of *my* country, the deep concern we have for those who suffer and those who sorrow. *I* have taken an oath today in the presence of God and *my* countrymen to uphold and defend the Constitution of the United States. To that oath *I* now add this sacred commitment: *I* shall consecrate *my* Office, *my* energies, and all the wisdom *I* can summon to the cause of peace among nations." R. Nixon, first Inaugural Address, Jan. 20th, 1969.

The pronoun *I* is usually more frequent in the SOTU addresses (mean: 1.3%) than in the inaugural allocutions (mean: 0.9%). Thus the over-use of the *Self* category in this passage extracted from his inaugural address is independent of the speech type. Moreover, in Nixon's addresses, the verbs are conjugated more frequently in the present tense (see previous example). Hart (1984) also points out that Nixon can be seen as a demagogue, being the president using familiar terms more frequently, using a limited vocabulary and promoting *Symbolism* (3.2% in Table 2 compared to 1.8% for Kennedy or 2.4% for Johnson).

With Reagan, America discovers the Great Communicator (but not a great orator or a great style master) (Hart, 1984). This president fits perfectly on television, accompanying his speeches with a voice and physical presence which provides an undeniable emotional embellishment. Seen as honest, sincere, and believing in simple values, the president knows how to adapt his speech to the context. Using fewer adjectives than his predecessors (see Table 1), he will emphasize simple phrases ("*we've come to a turning point*" or "*we've tried to fight inflation*") using a rather limited lexicon (a low MATTR value of 39.7% compared to 41.5% for Kennedy). This familiar vocabulary includes a large proportion of terms related to *Human* (e.g., *child, parent*). For this category, Table 2 indicates a percentage of 7.5% for Reagan compared to 6.4% for Nixon, of 5.7% for Kennedy. In addition, Reagan's addresses contain more symbolic (e.g., *freedom, America*), and religious expressions (e.g. *temple*) as shown by this example.

> "If *we* do that, if *we* care what *our children* and *our children's children will say* of *us*, if *we* want them one day to be thankful for what *we* did here in these *temples* of *freedom, we will* work *together* to make *America* better for *our* having been here, not just in this year or this *decade* but in the next century and beyond." R. Reagan, *State of the Union*, Jan. 25th, 1983.

One of Reagan's features is the higher proportion of verbs (in Table 1, 14.9% for Reagan compared to 12.8% for Kennedy), and words indicating *Action* (e.g., *achieve, deliver, recommend, teach*). With Reagan's presidency, references to God, or religion in general, become significantly more frequent, a rhetoric aspect followed by his successors as illustrated in the following passage:

> "I ask you to bow your heads. *Heavenly* Father, we bow our heads and thank *You* for *Your* love. … Make us strong to do *Your* work, *willing* to heed and hear *Your* will, and write on our hearts these words: "Use power to help people." For we are given power not to *advance* our own *purposes*, nor to make a great show in the world, nor a name. There is but one just use of power, and it is to serve people. Help us *remember, Lord. Amen*." G. H. Bush, Inaugural address, Jan. 20th, 1989.

In fact, Reagan is the president using the term *God* most often with a relative frequency of 1.15‰, followed by G. H. Bush (0.57‰), G. W. Bush (0.49‰), and Obama

(0.47‰). This reference introduced more frequently under Reagan's presidency is now part of the vocabulary of contemporary presidents.

According to Hart et al. (2013), the language of G. H. Bush's presidency is accompanied by a greater emphasis on patriotism, religious language (see example shown above), and references to citizens and people. The rhetoric is mainly assertive with an absence of doubt. The president employs the pronoun *I* frequently (*Self*: 2% compared to a mean of 1.2%) and presents the highest value in the category *Cognitive* (21% in Table 2). We can observe a slightly larger proportion of verbs (15.5%), a grammatical category that will increase with the next presidents (as indicated in Table 1). Overall, G. H. Bush's presidency is also characterized by a low percentage of big words (25.6% compared to a mean of 31.5%, see Table A.2), and a very short mean sentence length (18.8 tokens/sentence), the smallest value over all US presidencies (see Figure 2).

With Clinton, America knows the birth of the digital economy, but also a president who is faced with an impeachment proceeding (Lewinsky scandal). Clinton remains however one of the most popular presidents with a rhetoric combining a realistic tone (colloquial, concrete, with an interest in the *Human* (8.8% in Table 2 compared to a mean of 7.6%)), that avoids complex formulas. The MATTR indicates one of the lowest values (36.6%) since 1945, signaling a limited vocabulary and a clear tendency to repeat the same formulations and terms. For Americans, the president speaks a language they understand, he is one of them and they gain a sense of confidence in him (despite his lies in the Lewinsky affair) (Hart et al., 2013). The following passage indicates a high percentage of terms related to the Collectives thematic (e.g., *country*, *family*, *economy*):

> "The *world economy*, the *world* environment, the *world* AIDS crisis, the *world* arms *race*: they affect us all. Today, as an older order passes, the new *world* is more free but less stable. Communism's collapse has called forth old animosities and new dangers. Clearly, America must continue to lead the *world* we did so much to make." B. Clinton, first Inaugural Address, Jan. 20th, 1993.

The arrival of G. W. Bush marks the advent of a much more partisan presidency. The presidency wants to closely monitor the political agenda, conceives in secret the needed policies, and then sells them with enough authoritarianism and largely ignores the press (Jacobson, 2008). The presidential picture is created around the adjectives "*arrogant, critical, messianic*". Table 2 clearly indicates that this presidency can be characterized by using more emotional words, both positive (5.2%) and negative (3%). In the following passage, words related to the emotion categories are indicated in italics:

> "We will build our *defenses* beyond *challenge*, lest *weakness* invite *challenge*. We will *confront weapons* of mass *destruction*, so that a new century is spared new *horrors*. The *enemies* of *liberty* and our country should make no *mistake*: America remains *engaged* in the world, by history and by choice, shaping a balance of power that *favors freedom*. We will defend our allies and our *interests*. We will show purpose without arrogance. We will meet *aggression* and *bad faith* with *resolve* and *strength*." G. W. Bush, first Inaugural address, Jan. 20th, 2001.

During the 2004 re-election campaign, Kerry's and Bush's programs prove to be close (social security, global warming, embezzlement of large companies) (Slatcher et al., 2007). Overall, Bush's rhetoric is marked by an optimism and the frequent use of symbolic terms (3.1% in Table 2).

The election of Obama was the first to be marked by a wide use of the Internet and the social networks (Facebook, YouTube, blogs, Twitter). The first years passed in a difficult context (financial crisis and unemployment). The presidential tone, however, remains optimistic and the president insists on some key issues by using many repeats. The recourse to *Tentative* (1.8%), *Exclusive* (2.6%), and *Negation* (1.5%) are frequent, and these three categories present the highest values over the nine presidencies. He also chooses to return to a rhetoric emphasizing the *Concreteness* (5.4% as reported in Table 2) and *Human* (7.9%) terms. The tone is however less emotional than his predecessors (both categories *Posemo* and *Negemo* are below the mean, see Table 2). The percentage of big words (26.3% indicated in Table A.2) is low indicating a real concern to avoid complex formulations. Obama's rhetoric is also characterized by a higher frequency of story-telling to concretely illustrate a number (unemployment rate) or an action. Both are shown in the following example:

> "Today in America, a teacher spent extra time with a student who needed it, and did her part to lift America's graduation rate to its highest level in more than three decades. An entrepreneur flipped on the lights in her tech startup, and did her part to add to the more than 8 million new jobs our businesses have created over the past four years." B. Obama, *State of the Union*, Jan. 28th, 2014.

The 2016 US presidential election was characterized by two figures, H. Clinton & D. Trump, both unloved by the majority of Americans. Ignoring every norm of American politics and hoping to reflect the silent majority, Trump says what he thinks, and thus appears sincere and authentic. His rhetoric is centered around the high values for the categories *Inclusive* (11%, see Table 2, e.g., *together*, *with*) and *Symbolism* (3.4%) terms (e.g., *America*, *country*, *freedom*). In order to be understood by everybody, the mean sentence length is rather short (20.5 token/sentence, see Table A.2), the second lowest value over all US presidencies. Moreover, a high percentage of *Concreteness* terms (6.1%) boosts a direct formulation. Finally, this presidency shows the highest value in the categories *Social* (12.4%) and *Human* (9.2%). The following example illustrates these two aspects with words depicted in italics.

> "But to create this future, *we* must work with, not against, the *men* and *women* of law enforcement. *We* must build bridges of cooperation and trust, not drive the wedge of disunity and division. Police and sheriffs are members of *our community*. They are *friends* and *neighbors*, *they* are *mothers* and *fathers*, *sons* and *daughters*, and *they* leave behind *loved* ones every day *who* worry whether or not *they*'ll come home safe and sound. *We* must support the incredible *men* and *women* of law enforcement. And *we* must support the victims of crime." D. Trump, *State of the Union* Address, Feb. 28th, 2017.

8. Conclusion

The eloquence of the president, his power of persuasion (Neustadt, 1990) and, in general, his rhetoric and stylistic choices allow him to explain his positions and to justify his actions. For the United States before T. Roosevelt (1901-1908), the presidency does not fully match that vision. Simply by looking at the number of presidential speeches per year (Tulis, 1987), the president appears very infrequently in public to address his remarks, that are often limited to thanks. With the emergence of a strong presidential power (Nye, 2013), governmental speeches become more frequent and important.

By analyzing the general evolution of the presidential rhetoric over the past two hundred years, some of our measurements indicate a trend towards simplification. The sentences become shorter, the percentage of big words decreases (see Figure 3), and complex reasoning disappears. For other measures such as the MATTR or the lexical density (see Figure 3) do not corroborate such a simplification. The last presidents tend to employ more verbs and pronouns, while nouns and adjectives are reduced (see Table 1). The vocabulary is opened to a more poetic tone as well as presenting more abstract expressions and includes more religious vocabulary (Lim, 2002) (see examples in Section 7). The presidents offer an optimistic vision of the future, while being themselves more assertive; doubt tends to disappear from their addresses. The emotional terms tend to be more frequent and references to family and human beings occur more frequently (see Table 2). While being mostly enthusiastic, they tend to establish a dialogue, or, at least, a relationship with citizens and the people. The language style aims to be more intimate with a greater frequency of pronouns like *we*, or *I* (see Figure 1). The adoption of story-telling reinforces this tendency.

There is however a difference between what the president says and what he does or what he gets. Indeed, the essential purpose of the SOTU address is to explain the intention and legislative propositions of the White House and to justify budget requests to the Congress. On this last point, for the period from 1965 to 2002, the success rate of demands for credit by the president was 52% during his first term, and 39% during his second term (Hoffman & Howard, 2006). This finding indicates that if the presidential speech is the vehicle of the government's priorities, and a prerequisite for its actions, it is far from being imperial...

References

Arnold, T., & Tilton, L. (2015). *Humanities Data in R. Exploring Networks, Geospatial Data, Images, and Text*. Heidelberg: Springer-Verlag Press.

Baayen, H.R. (2008). *Analyzing Linguistic Data. A Practical Introduction Using R*. Cambridge: Cambridge University Press.

Biber, D., & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Leech, G. (2002). *The Longman Student Grammar of Spoken and Written English*. London: Longman.

Burrows, J.F. (2002). Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17, 267-287.

Caesar, J.W., Thurow, G.E., Tulis, J., & Bessette, J.M. (1981). The Rise of Rhetorical Presidency. *Presidential Studies Quarterly*, 11, 2, 158-171.

Carpenter, R.H., & Seltzer, R.V. (1970). On Nixon's Kennedy Style. *Speaker and Gavel*, 7, 41.

Gascuel, O., & Steel, M. (2006). Neighbor-Joining Revealed. *Molecular Biology and Evolution*, 23, 1997-2000.

Gelderman, C. (1997). *All the Presidents' Words. The Bully Pulpit and the Creation of the Virtual Presidency*. New York: Walker & Co.

Conover, W.J. (1971). *Practical Nonparametric Statistics*. New York: John Wiley & Sons.

Greenstein, F.I. (2009). *The Presidential Difference: The Leadership Style from FDR to Barack Obama*. Princeton: Princeton University Press.

Grzybek, P., Kelih, E., & Stadlober, E. (2008). The Relationship between Word Length and Sentence Length: An Intra-Systemic Perspective in the Core Data Structure. *Glottometrics*, 16, 111-121.

Hart, R.P. (1984). *Verbal Style and the Presidency*. Orlando: Academic Press.

Hart, R.P., Childers, J.P., & Lind, C.J. (2013). *Political Tone. How Leaders Talk & Why*. Chicago: Chicago University Press.

Hoffman, D.R., & Howard, A.D. (2006). *Addressing the* State of the Union. *The Evolution and Impact of the President's Big Speech*. Boulder: Lynne Rienner.

Humes, J.C. (1997). *Confessions of a White House Ghostwriter: Five Presidents and Other Political Adventures*. Washington: Regnecy Publ.

Jacobson, G.C. (2008). *A Divider, not a Uniter: George W. Bush and the American People*. New York: Pearson-Longman.

Jockers, M.L. (2014). *Text Analysis with R for Students of Literature*. Heidelberg: Spinger-Verlag.

Kalb, D., & Peters, G. (2007). State of the Union. *Presidential Rhetoric from Woodrow Wilson to George W. Bush*. Washington CQ Press.

Kolakowski, M., & Neale T.H. (2006). The President's *State of the Union* Message: Frequently Asked Questions. *Congressional Research Service*, n0 RS20021.

Labbé, D. (2007). Experiments on Authorship Attribution by Intertextual Distance in English. *Journal of Quantitative Linguistics*, 14, 1, 33-80.

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

Labbé, D., & Monière, D. (2008). *Les mots qui nous gouvernent. Le discours des premiers ministres québécois : 1960-2005*. Montreal : Monière-Wollank.

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

Lee, J.J., Cho, H.Y., and Park, H.R. (1999). N-gram-Based Indexing for Korean Text Retrieval. *Information Processing & Management*, 35, 427-441

Lim, E.T. (2002). Five Trends in Presidential Rhetoric: An Analysis of Rhetoric from George Washington to Bill Clinton. *Presidential Studies Quarterly*, 32, 2, 328-348.

Manning, C.D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: The MIT Press.

Mimno, D. (2012). Computational Historiography: Data Mining in a Century of Classics Journals. *ACM Journal on Computing and Cultural Heritage*, 5, 1-19.

Mitchell, D. (2015). Type-Token Models: A Comparative Study. *Journal of Quantitative Linguistics*, 22, 1-21.

Muller, C. (1992), *Principes et méthodes de statistique lexicale*. Paris : Champion.

Neustadt, R.E. (1990). *The Presidential Power and the Modern Presidents. The Politics of Leadership from Roosevelt to Reagan*. New York Free Press.

Nye, J.S. (2013). *Presidential Leadership and the Creation of the American Era*. Princeton: Princeton University Press.

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

Pennebaker, J.W. (2009). What is "I" saying? (guest post). The Language Log. http://languagelog.ldc.upenn.edu/nll/?p=1651 (access Feb. 19th., 2016).

Pennebaker, J.W. (2011). *The Secret Life of Pronouns. What our Words Say About us*. New York: Bloomsbury Press.

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B. D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., & Vidya, M.N. (2009). *Word Frequency Studies.* Berlin: De Gruyter Mouton.

Ridings, W. J., & McIver, S. B. (1997). *Rating the Presidents: A Ranking of U.S. Leaders, from the Great and Honorable to the Dishonest and Incompetent*. Secaucus: Carol Publishing.

Rule, A., Cointet, J.-P., & Bearman, P.S. (2015). Lexical Shifts, Substantive Changes, and Continuity in State of the Union Discourse. 1790-2014. In: *Proceedings of the National Academy of Sciences*, 112, 35, 1-8.

Rzhrtsky, A., & Nei, M. (1993). A Simple Method for Estimating and Testing Minimum-Evolution Trees. *Molecular Biology and Evolution*, 10, 1073-1095.

Savoy, J. (2010). Lexical Analysis of US Political Speeches. *Journal of Quantitative Linguistics*. 17, 2, 123-141.

Savoy, J. (2014). Authorship Attribution using Political Speeches. In: *Proceedings Qualico*, June, Reprint in Tuzzi, A., Benešová, M., & Mačutek, J. (eds.), 2015, *Recent Contributions to Quantitative Linguistics*. Berlin: De Gruyter Mouton.

Savoy, J. (2015a). Comparative Evaluation of Selection Functions for Authorship Attribution. *Digital Scholarship in the Humanities*, 30, 2, 246-261.

Savoy, J. (2015b). Text Clustering: An Application with the *State of the Union* Addresses. *Journal of the American Society for Information Science & Technology*, 66, 8, 1645-1654.

Savoy, J. (2016). Text Representation Strategies: An Example with the *State of the Union* Addresses. *Journal of the American Society for Information Science & Technology*, 67(8), 1858-1870.

Sherrill, R. (1967). *The Accidental President*. New York: Grossman.

Shogan, C.J., & Neale, T.H. (2012). The President's *State of the Union* Address: Tradition, Function, and Policy Implications. *Congressional Research Service*, no 7-5700.

Slatcher, R.B., Chung, C.K., Pennebaker, J.W. & 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, 63-75.

Sproat, R. (1992). *Morphology and Computation*. Cambridge: The MIT Press.

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, 1, 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*, Edmonton (AL), 252-259.

Tulis, J. (1987). *The Rhetorical Presidency*. Princeton: Princeton University Press.

Appendix

Table A.1. List of 45 US Presidents with their number of Inaugural and SOTU speeches together with their political affiliation

#	Name	Inaugural	SOTU	From	To	Party
1	George Washington	2	8	1789	1797	Ind.
2	John Adams	1	4	1797	1801	F
3	Thomas Jefferson	2	8	1801	1809	D–R
4	James Madison	2	8	1809	1817	D–R
5	James Monroe	2	8	1817	1825	D–R
6	John Quincy Adams	1	4	1825	1829	N–R
7	Andrew Jackson	2	8	1829	1837	D
8	Martin Van Buren	1	4	1837	1841	D
9	William H. Harrison	1		1841	1841	Whig
10	John Tyler		4	1841	1845	D
11	James Polk	1	4	1845	1849	D
12	Zachary Taylor	1	1	1849	1850	Whig
13	Millard Fillmore		3	1850	1853	Whig
14	Franklin Pierce	1	4	1853	1857	D
15	James Buchanan	1	4	1857	1861	D
16	Abraham Lincoln	2	4	1861	1865	R
17	Andrew Johnson		4	1865	1869	D
18	Ulysses S. Grant	2	8	1869	1877	R
19	Rutherford B. Hayes	1	4	1877	1881	R
20	James A. Garfield	1		1881	1881	R
21	Chester A. Arthur		4	1881	1885	R
22	Grover Cleveland	1	4	1885	1889	D
23	Benjamin Harrison	1	4	1889	1893	R
24	Grover Cleveland	1	4	1893	1897	D
25	William McKinley	2	4	1897	1901	R
26	Theodore Roosevelt	1	8	1901	1909	R
27	William H. Taft	1	4	1909	1913	R
28	Woodrow Wilson	2	8	1913	1921	D
29	Warren Harding	1	2	1921	1923	R
30	Calvin Coolidge	1	6	1923	1929	R
31	Herbert Hoover	1	4	1929	1933	R
32	Franklin D.	4	12	1933	1945	D
33	Harry S. Truman	1	7	1945	1953	D
34	Dwight D.	2	9	1953	1961	R
35	John F. Kennedy	1	3	1961	1963	D
36	Lyndon B. Johnson	1	6	1963	1969	D
37	Richard Nixon	2	5	1969	1974	R
38	Gerald R. Ford		3	1974	1977	R
39	Jimmy Carter	1	3	1977	1981	D
40	Ronald Reagan	2	7	1981	1989	R
41	George H. Bush	1	4	1989	1993	R
42	Bill Clinton	2	8	1993	2001	D
43	George W. Bush	2	8	2001	2009	R
44	Barack Obama	2	8	2009	2017	D
45	Donald Trump	1	1	2017	-	R

Ind.: Independent D-R: Democratic-Republican N-R: National-Republican
D: Democratic R: Republican

Table A.2. Overall stylistic measurements for some selected presidencies

	Wash.	Madis.	Wilson	Eisen.	Reagan	H. Bush	Obama	Trump	**Mean**
MSL	42.2	**48.3**	33.6	23.4	22.6	*18.8*	21.1	20.5	33.3
BW	32.9%	32.9%	27.9%	**36.1**%	28.7%	*25.6*%	26.3%	29.4%	31.5%
LD	42.4%	43.1%	*41.5*%	**50.3**%	47.3%	45.7%	46.4%	47.6%	44.8%
MATTR	**41.2**%	40.0%	*36.8*%	41.1%	39.7%	37.7%	38.9%	40.3%	38.3%

MSL: mean sentence length
BW: percentage of Big Words
LD: Lexical Density
MATTR: Moving Average Type-Token Ratio