Text Clustering: An Application with the *State of the Union* Addresses

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

This paper describes a clustering and authorship attribution study over the State of the Union addresses from 1790 to 2014 (224 speeches delivered by 41 presidents). To define the style of each presidency, we have applied a principal component analysis (PCA) based on the part-of-speech (POS) frequencies. From Roosevelt (1934), each president tends to own a distinctive style whereas previous presidents tend usually to share some stylistic aspects with others. Applying an automatic classification based on the frequencies of all content-bearing word-types we show that chronology tends to play a central role in forming clusters, a factor that is more important than political affiliation. Using the 300 most frequent word-types, we generate another clustering representation based on the style of each president. This second view shares similarities with the first one, but usually with more numerous and smaller clusters. Finally, an authorship attribution approach for each speech can reach a success rate of around 95.7% under some constraints. When an incorrect assignment is detected, the proposed author often belongs to the same party and has lived during roughly the same time period as the presumed author. A deeper analysis of some incorrect assignments reveals interesting reasons justifying difficult attributions.

Introduction

With the current technology, accessing a huge number of documents is no longer a real challenge. For example, Google *Ngram Viewer* (Michel et al., 2011) allows us to access around 4% of all printed books from 1800 by sending a query (a word or a list of words). With such a tool, we can observe the relative frequencies of various terms across the interval as, for example, the increasing use of some technologies (e.g., phone, computer) or the decrease of others (e.g., steam engine, telegraph). Linguistics can see the variation in usage of some synonyms such as *radio* and *wireless*. Using this website, Juola (2013) demonstrates that we can measure quantitatively the increasing complexity of Western culture.

Instead of being limited to comparisons of frequencies, Moretti (2005) and Jockers (2013) suggest that we can apply or generate more powerful models and tools to derive pertinent and synthetic information from text corpora. Based on word usage and metadata information, we can extract trees, maps, and graphs to generate and explore new facets in literary studies. Following this perspective, we want to automatically analyze a corpus in another domain (political science) and use synthetic representation to reveal the relationships between US presidents based on the content and writing style of their addresses.

Wishing to work with textual data of high quality, having none or a few spelling errors, we have selected political speeches possessing other advantages as well. Such documents are easy to access, without a fee or strict copyright. They can also cover a rather long timespan. Finally, they are usually relatively easy to read and interpret unlike some scientific documentation.

In this study, we have chosen a corpus reflecting US politics and history by selecting the *State of the Union* addresses. Based on 224 speeches delivered by 41 presidents, we want to analyze the similarities between presidents on two main dimensions. The first is related to the content of their addresses and the second reflects the style of the various tenants of the White House.

When analyzing the content of this corpus, we may assume that close relationships can be detected between presidents belonging to the same political party. According to this first hypothesis, we can expect to have two large clusters, one Democrat and one Republican, at least for the last century. In fact during a Democratic presidency, we may expect more topics about education, family, welfare, and healthcare. Under a Republican administration, subjects related to free enterprise and business, reduction of

Received January 17, 2014; revised February 27, 2014; accepted March 3, 2014

 DOI: 10.1002/asi.23283

expenses, or a stronger support for the military sector should be more frequent (Petrocik, Benoit, & Hansen, 2003).

As an alternative hypothesis, we can imagine that each Commander-in-Chief has his own priorities, different from his predecessor's. The new president will impose (or will try to impose) his own political agenda, establishing the issues he wants to address or asking the Congress to discuss them. Following this hypothesis, each chief of the Executive will have a distinct vocabulary.

We also want to analyze the style of these speeches. In this case, we can postulate that the time period will have an impact on the expressions and formulations used by the presidents. Current presidents will not use the same style as that used by the Founding Fathers. Thus can we detect only a few distinct presidential styles or as many as the number of presidents?

The rest of this paper is organized as follows. The second section, Related Work, presents an overview of related topics while the third section describes the corpus used in our experiments. Text Clustering Based on Topical Features uses the Part-Of-Speech (POS) to derive similarities between the president's styles. Text Clustering Based on Stylistic Considerations exposes an intertextual distance measure and applies it to the *State of the Union* addresses based on their content. Authorship Attribution explains the results of a clustering method based on the style of the different presidents. Finally, the Conclusion describes and comments about an experiment on authorship attribution using all *State of the Union* addresses.

Related Work

In order to automatically generate a synthetic view from a given corpus, there is the Google Ngram Viewer (Michel et al., 2011) which provides access to around 4% of all printed books. With such a tool, we can compare the relative notoriety of different persons (e.g., Roosevelt, Lincoln, and Washington) from 1800 until 2008 based on their relative annual frequencies. This tool, however, does not directly generate a synthetic view, but allows the user to explore the sources according to his requests.

To automatically extract a set of topics from a corpus, we can opt for the latent Dirichlet Allocation (LDA) or topic model proposed by Blei, Ng, and Jordan (2003), (Blei & Lafferty, 2009). In this context, the corpus is considered as generated by a probabilistic model. More precisely, each document is viewed as containing one or more topics, and we model each document as a distribution over this set of topics. Each topic is represented as a distribution over all words. The word order is irrelevant inside each document. Given a corpus and a number of topics, the LDA returns, for each topic, a list of words with their probability of occurrence and, for each document, a probability for each topic. The output is therefore not a synthetic view but a complete probabilistic description of the underlying corpus.

The set of experiments on literary history described by Jockers (2013) seems more closely related to our investigation. In this case, the author presents a set of available techniques to analyze the corpora of literary works as, for example, the word clouds or the LDA techniques for content analysis, the distribution of genre over time, and the distinction between genders according to the vocabulary used.

More closely related to the political discourse analysis, various studies have been conducted to examine the electoral speeches to discover the recurrent topics or specific themes of each candidate. However, such analyses do not directly represent the topics, but rather, they focus on the vocabulary used and the occurrence frequency of the most specific terms. In this vein, we can mention a lexical study of the French presidential election in 2007 (Calvet & Véronis, 2008) and the US campaign in 2008 (Savoy, 2010).

Finally, we can mention the works of Labbé and Monière (2003, 2008) covering a relatively long period of governmental speeches (1945–2000). Written in French, these two researchers compare three parliamentary systems, by analyzing the *Speeches of the Throne* (Canada), the *Inaugural Addresses* (Quebec), and general policy statements (France). Based on the vocabulary used by the different governments, these studies show how the content of the speeches evolved during the last 50 years. Moreover, the similarities between speeches written by governments coming from different parties are greater than expected. The institutions and the current issues tend to impose a similar content, even for Prime Ministers coming from different affiliations. A similar conclusion was found for Italy over the last 50 years (Pauli & Tuzzi, 2009).

The State of the Union Addresses

The choice of the *State of the Union* addresses as a political corpus can be explained by the following reasons. First, according to the US Constitution (Article II, Section 3), the president must provide information to the Congress about the state of the Union and "*measures as he shall judge necessary and expedient.*" Such speeches provide both a picture of the current situation, indicate the president's priorities and the proposed legislative agenda. Second, as the US plays the role of leader, the US president is viewed as the most powerful person in the world. His decisions often have a worldwide impact. In this perspective, the *State of the Union* addresses have an interest not limited to a single country. Third, these speeches cover a relatively long timespan starting in 1790 and covering more than two centuries. Fourth, some of them are well-known for defining a political position held for decades such as the Monroe Doctrine (1823), the *Four Freedoms* (Roosevelt in 1941), or the *War on Poverty* (Johnson in 1964). In some speeches, we find the first occurrence of well-known expressions such as the *axis of evil* (Bush in 2002). A more detailed analysis of the form and political functions of these presidential messages can be found in Kolakowski and Neale (2006), and Shogan and Neale (2012).

DOI: 10.1002/asi

To create the corpus, we downloaded all the addresses from the website http://www.presidency.ucsb.edu. The corpus contains 224 speeches delivered by 41 US presidents. The first address was delivered by G. Washington (January, 8th, 1790) and the last by B. Obama (January, 28th, 2014). For two presidents (W.H. Harrison [1841] and J.A. Garfield [1881]), we do not have any *State of the Union* addresses because their term was too short (a few months). We have also removed the single address given by Taylor (1849) and the message given by Truman in 1946. This remark is five times longer than the others on the one hand and, on the other, it is the only one delivered solely in written form. For the same reasons, we also ignore the message delivered by Carter in 1981. Cleveland appears twice as president (1885–1888 and in 1893–1896) corresponding to his two terms interrupted by B. Harrison's presidency (1889–1892). Starting with 1853, we count 18 Republican (R) presidencies, and 13 Democratic (D). A more complete list can be found in the Appendix.

For each speech, we added a few meta-tags to store document information (e.g., date, author), and we also cleaned them up by replacing certain UTF-8 coding system punctuation marks with their corresponding ASCII code symbols. This involved replacing single (‘’) or double quotation marks (“”), with their ASCII equivalents and the removal of diacritics found in certain words (e.g., *naïve*). Moreover, the contracted forms have been replaced by their equivalent full forms (e.g., *don't* into *do not*).

To represent each speech, we can use the word-tokens (or surface words, or simply tokens) (e.g., *is, were, been* or *armies, army*) or the word-types (or lemma, entry in the dictionary). In the latter case, various word-tokens belonging to the same root are regrouped (e.g., *be* or *army* in our previous example). In the current study, we retained the word-types in order to ignore the possible variations due to syntax. Thus, we do not consider the two word-types *I* and *me* as dissimilar and thus we merged them under the common headword *I*. We considered the distinction between the two grammatical cases (*I,* subject or nominative case vs. *me* direct object or in the accusative case) to be of secondary importance, and thus decided to group both word-tokens under the same word-type. We also applied the same conflation to the other pronouns (*we* or *us*, *they* or *them*, *he* or *him* and *she* or *her*).

To define the corresponding work-type to each token, we used the POS tagger developed by Toutanova, Klein, Manning, and Singer (2003). 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-third-person singular present, VBZ third-person singular present), adjectives (JJ, JJR adjective in comparative form), personal pronouns (PRP), prepositions (IN), and adverbs (RB). These morphological tags (Marcus, Santorini, & Marcinkiewicz, 1993) correspond mainly to those used in the Brown corpus (Francis & Kučera, 1982). With this information we were then able to derive the word-type 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).

After this tagging, our US corpus contains 1,955,699 word-tokens for 20,589 distinct word-types (length of the vocabulary). When considering the occurrence frequency, we have 6,242 *hapax legomena* (word-types appearing only once, and corresponding to 30.3% of the whole vocabulary) and 2,426 *dis legomena* (word-types occurring exactly twice, representing 11.8% of the vocabulary). The definite determiner (*the*, 151,068 occurrences) is the most frequent word-type, followed by *of* (97,818), the comma (96,128), *be* (65,455), the full stop (61,563), *to* (60,182), *and* (59,920), and *in* (38,335).

Analysis of the speeches gives the mean length as 8,725 word-tokens (standard deviation: 5,847.5). The longest remarks were delivered by Taft in 1910 (30,773 tokens) and the shortest by Washington in January 1790 (1,180 tokens). When considering the mean length per president, Adams (1797–1800) wrote the shortest speeches (average of 1931 word-tokens per speech) while Taft (1909–1912) is the author, in mean, of the longest addresses (24,655 word-tokens).

Finally, in our corpus, when two presidents have the same family name, we must be able to distinguish between the two persons. Therefore, we denote *HBush* for the father (George H. W. Bush) and simply *Bush* for his son (George W. Bush). The name *Roosevelt* is reserved for Franklin D. Roosevelt (1934–1945) and by *TRoosevelt* we mean Theodore Roosevelt (1901–1908). The name *Johnson* signals Lyndon B. Johnson (1964–1969) while *AJohnson* corresponds to Andrew Johnson (1865–1868). The name *Adams* designates John Adams (1797–1800) while his son is indicated by *QAdams* (John Quincy Adams, 1825–1828).

Part-of-Speech Analysis

As a possible feature set to discriminate between the styles of the US presidents, we can consider the distribution of the different POS categories. To achieve this, we form a profile for each president composed of all his speeches. To represent them, we consider only the POS tags, including also the full stop (period), and we regroup all other punctuation symbols. With these two elements, we may detect a text showing long sentences with a lower rate of periods and a higher rate of other punctuation marks (e.g., comma).

Based on this information, we can discriminate between the presidents, verifying whether they all use a distinct style to present their *State of the Union* addresses or whether they opt for the same (or very similar) form, under the

DOI: 10.1002/asi

assumption that the formal context will impose such a strict norm.

Based on the occurrence rate of each POS, Obama is the US president using verbs (16% of his tokens) and adverbs (9.4%) most frequently. This aspect indicates a speech oriented more towards action. The highest rate of nouns can be found with Hoover (1929–1932) (22.8%), specifying that the author tries mainly to explain the situation (e.g., the economic downturn in the 1930s). McKinley (1897–1900) uses names at the highest rate (5.9%). Eisenhower opts clearly for adjectives (9.2%), while Clinton prefers pronouns (8.9%). The dollar sign ($) appears more frequently under Kennedy (2.7%) as well as punctuation marks other than the full stop (6.6%). This high rate signals a preference for long sentences (with an abundance of commas). At the opposite end, *HBush* (the father) employs the full stop most frequently (5.2%), indicating a bias in favor of short sentences.

Instead of limiting the analysis on each category separately, we can position each president according to their occurrence frequencies of verbs (*Verb*), modal verbs (*Modal*), adverbs (*Adverb*), adjectives (*Adj*), nouns (*Noun*), names (*Name*), pronouns (*Pronoun*), determiners (*Det*), prepositions (*Prep*), conjunctions (*Conj*), numbers (*Number*), dollar signs (*$*), periods (*Dot*), and other punctuation marks (*Punct*). To achieve this visual representation, we opt for the principal component analysis (PCA) (Lebart, Salem, & Berry, 1998; Baayen, 2008) depicted in Figure 1.

In this figure, the horizontal axis indicates the opposition between the frequent use of determiners and prepositions on the left, and pronouns, modal forms, and adverbs shown in the right part. The vertical axis signals the frequent use of nouns, adjectives, and numbers (upwards direction) whereas verbs are associated with the downward direction. Of course, the exact position of each president is given by the 14 different categories, a number of dimensions too high to be represented exactly in a two dimensional paper. Thus in Figure 1, the PCA generates two orthogonal composite components taking into account 40.3% + 15.9% = 56.3% of the total underlying variability.

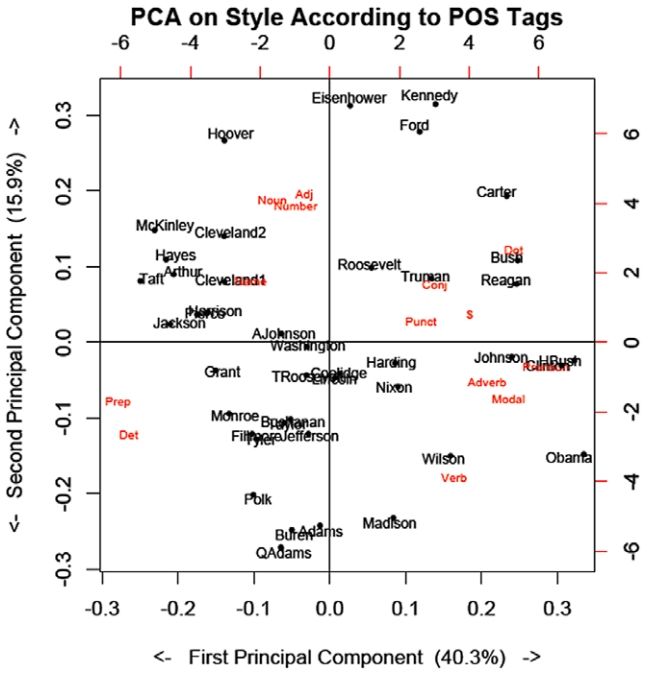

FIG. 1. Representation of each US president according to their usage of different Parts-of-Speech. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

In the center of Figure 1, where the two axes are crossing, we encounter presidents having an average use of all POS categories, such as Andrew Johnson (*AJohnson*, 1865–1868), Washington (1790–1796), Coolidge (1923–1928), and not too far, T. Roosevelt (1901–1908), and Lincoln (1861–1864). On the bottom of the figure, we have mainly the first presidents (Adams (1797–1800), QAdams (1825–1828), Madison (1809–1816)) together with Polk (1845–1848), and Van Buren (1837–1840). On the left, we can find mainly presidencies covering the end of the 19th century and the beginning of the 20th century such as Taft (1909–1912), Arthur (1881–1884), Hayes (1877–1880), Harrison (1889–1992), Cleveland1 (1885–1888), Cleveland2 (1893–1896), McKinley (1897–1900), and as an exception Jackson (1829–1836). From Roosevelt (1934) (upper right), the different presidents tend to appear in the top right section of the figure, using less determiners and prepositions. Moreover, instead of appearing concentrated in a small region (like the majority of the Founding Fathers), each of them tends to adopt a distinctive style. We can observe that Eisenhower, Kennedy, Ford, and Carter employ nouns, adjectives, and numbers more frequently, whereas Obama, *HBush* (father), Clinton, or Johnson have a bias in favor of pronouns, adverbs, and modal verbs.

Text Clustering Based on Topical Features

In order to regroup similar texts together, we need to define an intertextual similarity or a distance measure between two speeches or two groups of speeches (profile). When opting for a distance measure, a small value indicates that the two texts are very similar and share many characteristics. On the other hand, a large value signals very dissimilar documents having only a few properties in common.

In this study we use the intertextual distance proposed by Labbé (2007) returning a value between 0 and 1 depending on the degree of overlapping between the two texts. A value of 0 indicates that the two texts are identical, using the same vocabulary with the same frequencies for all terms. A distance of 1 specifies that the two speeches have nothing in common (e.g., they are written in two different languages, with no word having the same spelling). Between these two limits, the returned value depends on the number of words appearing in both texts and their occurrence frequencies.

More formally, the distance between the Text A and B (denoted D(A,B)) is given by Equation (1) where n_A indicates the length (number of tokens) of Text A, and tf_{iA} denotes the

DOI: 10.1002/asi

(absolute) term frequency of word-type i (for $i = 1, 2, \ldots, m$) in the text A. The length of the vocabulary is indicated by m. Usually Text B does not have the same length (in our case, we assume that the length of Text B is larger than Text A). We need therefore to reduce the longest text by multiplying each of its term frequencies (tf_{iB}) by the ratio of the two lengths as indicated in the second part of Equation (1).

$$D(A,B) = \frac{\sum_{i=1}^{m} \left| tf_{iA} - \hat{tf}_{iB} \right|}{2 \cdot n_A} \quad \text{with} \quad \hat{tf}_{iB} = tf_{iB} \frac{n_A}{n_B} \quad \text{and with} \quad n_A = \sum_i^m tf_{iA} \tag{1}$$

Finally, to return valid measurements, the length difference between the two texts must be smaller than eight times, and each text must contain at least 5,000 words (in our corpus, Adams's profile is the smallest containing 7,725 word-tokens).

This intertextual measure is a distance measure respecting the following properties (Labbé, 2007). The distance to itself equals to 0, meaning that D(A,A) = 0. It is symmetric, and thus D(A,B) = D(B,A). Finally, this measure respects the triangle inequality with D(A,C) ≤ D(A,B) + D(B,C).

In our study, we need to select the vocabulary reflecting the targeted application. To define a measure revealing the topics presented in the speeches, we ignore the functional words (such as determiners, prepositions, conjunctions, pronouns, and auxiliary or modal verbs) having no clear and important meaning (e.g., *the*, *of*, *you*, *we*, *have*, *does*, etc.). In the current study, we define this list by removing the top 300 most frequent word-types. Moreover, word-types occurring just once or twice in the corpus will be removed (Manning & Schütze, 1999; Sebastiani, 2002). Such rare word-tokens tend to be marginal and correspond usually to names (locations or persons), very infrequent subjects or spelling errors.

We compute the intertextual distance based on the president's profile (concatenation of all his addresses). After applying this distance measure, we can return a symmetrical matrix composed of $41 \times 41 = 1{,}681$ values which does not represent a synthetic view of the topical relationships between the presidents.

To achieve this objective, we apply an automatic classification scheme (Kaufman & Rousseeuw, 1990; Lebart et al., 1998). The result is a dendrogram tree showing a set of clusters with similar profiles. This classical representation, based on the Ward method, is depicted in the Appendix. Recently, and mainly in genomic studies, such distance matrices can be represented by a tree-based visualization respecting approximately the real distances between all nodes (Bartélémy & Guénoche, 1991; Baayen, 2008; Paradis, 2011). We adopt this new representation and the result is displayed in Figure 2.

In this figure, the distance between two presidents is indicated by the length of the lines needed to connect them. For example, we start with the first president, follow the branch until we reach the backbone, go along the backbone and then select the track leading to the second person. In this figure, the longest distance (0.917) connects Obama and Adams. The second longest distance (0.912) can be found between Adams and Clinton. The two closest presidents are Jackson and his successor Van Buren (0.35), whereas the second shortest distance (0.362) joins Obama with Clinton.

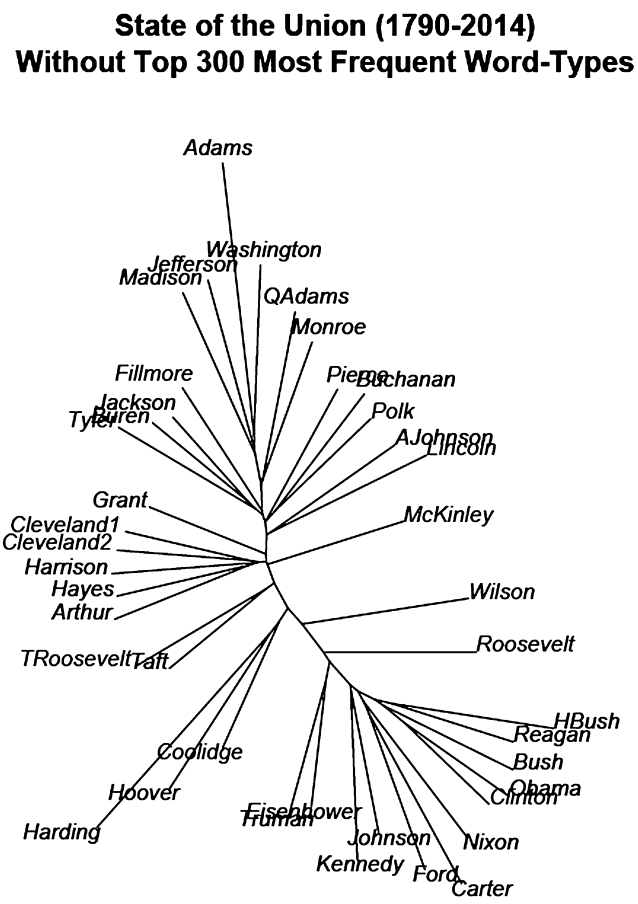

FIG. 2. Tree-based representation of the similarities between the profiles of the US presidents (topical word-terms).

In this figure, starting with Roosevelt (located on the right) and going clockwise, we can form the first cluster composed of the contemporary presidents (*HBush* (father) and Reagan closely related, Bush (son) and the tandem Obama–Clinton also strongly linked). Nixon, Carter, and Ford form the second group, and the third cluster is composed by Kennedy and Johnson. In the backbone, the distance between these three clusters is relatively short. Therefore these three clusters can be regrouped in a larger cluster at a higher level. With an increased distance, the tandem Eisenhower–Truman seems to make a bridge between the presidencies after 1960 and Roosevelt (1934–1945). This latter president is, like Wilson (1913–1920), isolated, clearly denoting a transient presidency that was faced with new questions and problems that required the development of new terminology and vocabulary.

DOI: 10.1002/asi

In the middle, three Republican presidents (Coolidge [1923–1928], Hoover [1929–1932], and Harding [1921–1922]) form a strongly related group located away from Roosevelt. Near the top, we see the Republican tandem Taft (1909–1912) and T. Roosevelt (1901–1908), which precedes a larger cluster formed by Cleveland1 (1885–1888), Cleveland2 (1893–1896), Harrison (1889–1892), Hayes (1877–1880), and Arthur (1881–1884). In this last group, we have only one Democrat (Cleveland) who appears twice (corresponding to his two terms). Above this group we see two isolated presidents, namely Grant (1869–1876), and in the opposite direction, McKinley (1897–1900). The duo A. Johnson (1865–1868) and Lincoln (1861–1864) starts an era of Republican Presidencies that will dominate US politics from 1861 until 1912 (with the exception of the two terms of Cleveland).

Moving upwards, we can see two Democrat clusters, the first composed of three presidents, namely Pierce (1853–1856), Buchanan (1857–1860), and Polk (1845–1848), and the second with Jackson (1829–1836), Van Buren (1837–1840), and Tyler (1841–1844). In the same time period, we also have Fillmore (1850–1852), belonging to the Whig party representing a parenthesis during a long sequence of Democrat tenants in the White House (from Jackson [1829] until Buchanan [1850]).

In the top part, the first six US presidents are subdivided into a group formed by the first four (Washington [1790–1796], Adams [1797–1800], Jefferson [1801–1808], and Madison [1809–1816]) and a duo formed with Monroe (1817–1824) and Quincy Adams (1825–1828).

This synthetic representation is based on the topical similarity between the presidents. When they are faced with the same problems and difficulties and propose similar responses, they tend to use the same vocabulary. Of course this assumption is not always satisfied because, when analyzing a given issue (e.g., immigration), one president may prefer describing it with abstract terms (e.g., *immigration*) whereas another may emphasize the persons involved in this question (e.g., *immigrants*). In general, however, the overlap is higher when discussing similar issues.

The general trend appearing behind Figure 2 is relatively clear. The time period tends to play the most important role in the relationships between presidents. The formation of clusters is strongly related to the timespan of each presidency. For each period, the president was not able to impose his own political agenda on the Congress without considering the current questions. Based on this finding, we can then describe the US history as subdivided according to four main epochs.

First, we can find the young Republic from Washington to Madison (1790–1824) followed by a transient period represented by Monroe and Quincy Adams. A second Democratic period covers the presidencies of Jackson to Buchanan (1829–1860), with Fillmore (1850–1852) as a free electron. Third, starting with Lincoln, we have a long Republican period until Hoover (from 1861 to 1932), a period starting with the Civil War and corresponding to the birth of an industrial nation. In this epoch, Grant and McKinley present a somewhat distinctive profile. Even if Wilson (1913–1920) belongs formally to this time period, the content of his speeches corresponds to a distinct presidency (Wilson was not the only Democrat in this period. We have also the two Cleveland presidencies). Fourth we have the last 10 presidents from Kennedy to Obama (1961–2014) who are preceded by a transient phase with Roosevelt to Eisenhower (1934–1960).

As we can see from this analysis, party affiliation is not the main explanation of the relationships between presidents, but its influence is not insignificant. For example, when inspecting the last 50 years, party affiliation can explain the generation of pairs such as Obama–Clinton, *HBush* (father)–Reagan, or Kennedy–Johnson, as well as the trio Harding, Coolidge, and Hoover. As a counter-example, we might mention the duo Truman (D) and Eisenhower (R)[1] or the rapprochement of Carter (D) towards the duo Ford–Nixon (R). But such clusters are less numerous.

Text Clustering Based on Stylistic Considerations

As an alternate view to detect relationships between presidents, we can ground the association according to each president's writing style. To achieve this, we may take account of the number of distinct word-types, the vocabulary richness, the sentence length, etc. (Baayen, 2008). In order to reduce the number of possible features, we can limit them to all pronouns that have been used to detect the psychological status of the author as well as to determine their gender (Pennebaker, 2011).

In the current study, we use only the k most frequent word-types (with $k = 300$) to compute the intertextual distance to reflect the stylistic information. The relative frequencies of these very frequent terms, mainly composed of functional words, tend to represent the fingerprint of each particular author and have been found effective in various authorship attribution studies (Burrows, 2002; Juola, 2006; Zhao & Zobel, 2007; Savoy, 2014).

As for the topical aspects, we compute the intertextual distance based on the president's profile (concatenation of all his addresses) using the $k = 300$ most frequent word-types. Then we can generate a tree-based representation as depicted in Figure 3. It is important to keep in mind that we have no term in common between the clusters shown in Figure 2 and 3. The top 300 most frequent word types have been ignored in Figure 2, and they are only used to generate Figure 3.

In this figure, the longest distance (0.337) connects Obama with Quincy Adams and the second longest (0.336) links Clinton with Quincy Adams. The third longest distance (0.331) can be found between Clinton and Madison. The two closest presidents are Jackson and his successor Van Buren (0.069) whereas the second shortest distance (0.072) joins

[1]If formally Eisenhower was a Republican, he entered only late 1951 in the political arena, and could have chosen to run for the Democrats.

DOI: 10.1002/asi

two Republicans Hayes and McKinley, and the third (0.075) presidency Cleverland1 and Cleverland2.

When comparing the style-based clustering (Figure 3) and the content-based view (Figure 2), we mainly see the same general pattern. The time period tends to have a clear impact on both clustering results even if the two features sets have nothing in common.

Instead of describing the clustering based on the stylistic elements, we will focus on the main differences between the two figures. Starting on the top of Figure 3, we see a similar configuration, with a stronger relationship between three duos, namely Obama–Clinton, Reagan–*HBush* (father), and Eisenhower–Truman. But from a stylistic point of view, Johnson is no longer closely related to Kennedy, and Carter is further away from both Nixon and Ford. If Roosevelt and Wilson present a distinct style, this particular aspect is also present for Harding (1921–1922), Coolidge (1923–1928), and Hoover (1929–1932). Therefore, we can conclude that these last three presidents are talking about the same questions (shown in Figure 2) but are using a distinctive style (depicted in Figure 3).

Taft (1909–1912), who was closely related to T. Roosevelt (1901–1908) in the content-based clustering, has a style closely related to a large cluster formed by Harrison (1889–1892), Hayes (1877–1880), Arthur (1881–1884), Grant (1869–1876), McKinley (1897–1900), and the two Cleveland presidencies (Cleveland1, 1885–1888, and Cleveland2, 1893–1896).

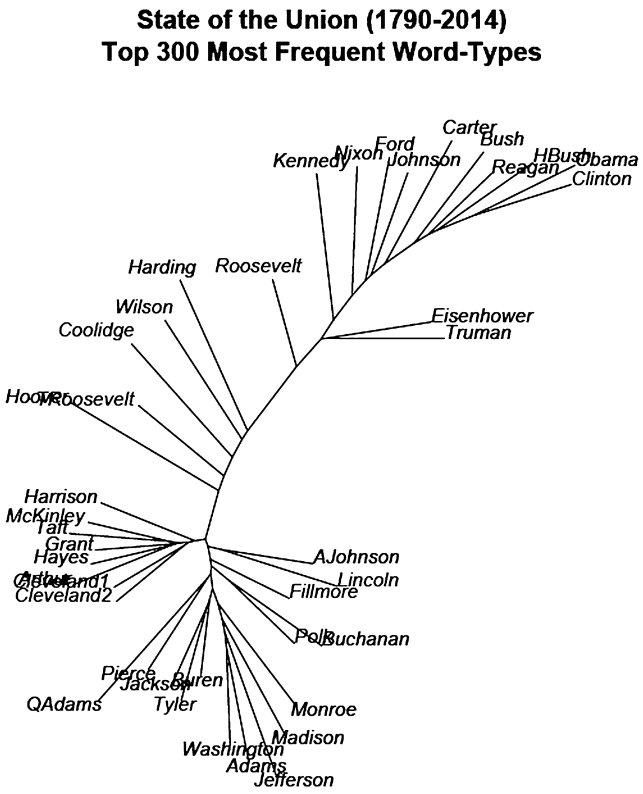

FIG. 3. Tree-based representation of the similarities between the profiles of the US presidents (style-based on the top 300 most frequent word-types).

The remaining clusters are very similar to those found in the content-based representation, with the duo Lincoln (1861–1864) and his successor A. Johnson (1865–1868), or with the trio of Jackson (1829–1836), Van Buren (1837–1840), and Tyler (1841–1844). The Democrat pair Polk (1845–1848) and Buchanan (1857–1860) is not directly related to Pierce (1853–1856), whose style is more related to Quincy Adams (1825–1828). The Founding Fathers group appears exactly in the same order in both the content and style-based clustering.

Authorship Attribution

Based on the relationships between presidents found either on the content of their speeches or their style, we could consider that merging the two sources of information could provide a strong signal to predict the presumed author behind a given address. In this case, we view the authorship attribution scheme based on the semantic content and the style of the presidential speeches.

To achieve this goal, we reuse the same intertextual distance (Labbé, 2007) but using all the available word-types. For each address, we can determine its possible author by computing its nearest neighbor (*k*-nearest neighbor, or *k*-nn) (Witten, Frank, & Hall, 2011). The author of this closest speech will be the most probable author of the query address.

Of course, this is not strictly an authorship attribution because we know that behind each well-known politician there is usually a speechwriter (or a team of ghost writers). For example, behind Kennedy we can find the name of Sorensen[2] (Carpenter & Seltzer, 1970), Favreau behind Obama, and even Madison and Hamilton behind some speeches delivered by Washington. Moreover, to take the latest events into account, the president might change some passages before delivering a message.

To automatically determine the real author with a high degree of credibility, the length of the disputed text must have 5,000 tokens or more (Labbé, 2007). This constraint is respected by 161 of the 225 speeches. However, many addresses are close to this limit (having between 4,500 to 4,800 tokens). Thus we will not take into account this first constraint.

The second constraint is more important and should be respected (Labbé, 2007). The assignment is reliable only when the distance between the two texts is small. The nearest neighbor approach will always detect a closest neighbor, even if the distance is rather long. In such cases, the proposed assignment is clearly less certain. We therefore need to define a limit under which the attribution will have a high degree of certainty, from an assignment based on a longer distance that can be interpreted as a simple indication without strong support.

[2]But Sorensen said "If a man in a high office speaks words which convey his principles and policies and ideas and he's willing to stand behind them and take whatever blame or therefore credit go with them, [the speech is] his."

DOI: 10.1002/asi

Having 224 speeches, we can compute the distance between all pairs of speeches resulting in a total of $([224 \times 224] - 224)/2 = 24{,}976$ distances. This amount is obtained by treating the distance between A and B is the same as the distance between B and A (symmetry). Moreover, the distance between A and A is always 0, and presents no interest.

We assume that the underlying distribution for these 24,976 values follows a normal distribution. In the current case, the sample mean is equal to 0.3886 (median: 0.392) with a standard deviation of 0.067. According to this assumption, we can determine a distance limit corresponding to 0.5% of all values which is $\lim_c = \text{mean} - 2.58 \times \text{standard deviation} = 0.2156$. All attributions based on a distance smaller than $\lim_c$ will be interpreted as *good evidence*.

We can also consider a less strict limit corresponding to 2.5% of all distance values with $\lim_p = \text{mean} - 1.96 \times \text{standard deviation} = 0.2571$. When the distance defining an attribution is smaller than $\lim_p$ but greater than $\lim_c$, we can interpret this assignment as *plausible*. Finally, an attribution based on a distance larger than $\lim_p$ must be viewed as *possible*, given without certainty.

When asking the system to attribute each of the 224 speeches to its author, the attribution scheme finds 45 speeches having a distance smaller than $\lim_c$ (and corresponding to *good evidence*). For all these cases, the proposed attribution is correct. In addition, we can find 71 cases where the assignment can be interpreted as *plausible* (the distance with the nearest neighbor is larger than $\lim_c$ but smaller than $\lim_p$). In this second set, the assignment is correct for 66 cases.

When taking into account the distance, the automatic attribution produces a success rate of $(45 + 66)/(45 + 71) = 95.7\%$. For the remaining 108 addresses, the system can only provide a possible assignment without any certainty. In this last set, we can detect 79 correct attributions.

Looking more carefully at the incorrect assignments, we can find interesting explanations. In 1964, the *State of the Union* address was delivered for the first time by Johnson. For the authorship attribution scheme, the most probable author is Kennedy. In fact President Kennedy was assassinated November 22nd, 1963, and the *State of the Union* address was delivered January 8th, 1964. Clearly the time was too short to have a new team of ghostwriters to write a completely new speech more closely reflecting Johnson's style and views. So we can assume that this *State of the Union* address was written by Kennedy's team.

As another example we can analyze the first speech delivered by G.H. Bush (father), February 9th, 1989. This address was attributed to Reagan (1982–1988) by the system. First, this was not really a *State of the Union* address but this speech was delivered to Congress and lays out the objectives for the new administration. Thus, like a *State of the Union* address, this speech is delivered in front of the Senate and the House of Representatives. Both the form and the content correspond clearly to a *State of the Union* address (and listed as it in the website). In this case, we see the influence of the previous administration (leaving January 20th, 1989) during the very first months of the new one.

A similar scenario appears with the first speech delivered by Bush (son) in 2001. In this case, the system indicates Clinton as the most probable author, reflecting the fact that the new presidency was faced with similar issues and difficulties as the previous one. As a second explanation, we might recall that this speech was the first one delivered before the attacks of September 11th, 2011. After this tragic event, the Bush administration focused more on the terrorist questions and homeland security and less on issues related to the first year of the presidency. For the system, the first speech is therefore distant from other Bush speeches, and closer to a Clinton address.

The hardest attribution problems can be found in the first 12 speeches, eight delivered by Washington (1790–1796), and four by Adams (1797–1800). In those cases, the system assigns correctly only three addresses to Washington, and none to Adams. The attribution scheme indicates Jackson (1829–1836) as the most probable author, a president sharing similar political views with Adams and Washington, and like the latter he was also an army general. The stylistic similarity between Jackson and Washington can be viewed on the middle of Figure 1 (based on POS information) where the two names appear relatively close together. Finally we must mention that the automatic attribution is less reliable when a disputed text is short. The mean length of these first twelve speeches is 2,153 word-tokens, while the mean over the 224 speeches is 8,725 word-tokens. Clearly, these first addresses are shorter than the mean, and thus more problematic to attribute with a high degree of certainty.

Conclusion

The corpus of the *State of the Union* addresses shows us the issues and difficulties facing the United States throughout its existence. Because the context and the content of these speeches are defined by the Constitution, they represent a fixed snapshot of the situation in the country on an annual basis. Written by the president (or his ghostwriter[s]), these addresses indicate also the legislative intents and priorities of the government.

Based on this corpus, we have established a tree-based representation depicting the relationships between the presidents. To achieve this, we represent each president by the word types occurring in all his speeches together with their frequencies. Based on the semantic content (ignoring both the top 300 most frequent word types or those appearing once or twice), the resulting figure indicates that time tends to play a major role in establishing connections between presidents. Presidencies appearing in the same timespan tend to appear in the same clusters. Clearly the society and culture in which Washington lived is distinct

DOI: 10.1002/asi

from the current one, meaning that Obama and Washington are faced with different issues. When studying the literature (Hughes, Foti, Krakauer, & Rockmore, 2012) or the linguistic evolution of language (Juola, 2003), other studies have also found that chronology plays an important role.

The party affiliation tends also to have an impact when establishing similarities between US presidents. This importance is however secondary and we do not see only two main groups, one Democrat and one Republican. When inspecting a short timespan (20 or 30 years), we often encounter two or three presidents strongly connected and belonging to the same political party. Thus in a relatively short period, political affiliation can explain the relationship between two presidencies (e.g., Clinton–Obama, Reagan–Bush [father], Jackson–Tyler–Van Buren, Polk–Buchanan).

Sing the top 300 most frequent word-types (composed mainly of functional words) to define the author style, the resulting clustering tends to show similar clusters to that of the content-based representation. One of the main differences is the presence of smaller clusters when considering style. Using the POS information, our stylistic analysis reveals that since Roosevelt (1934–1945), each president tends to adopt a style relatively distinctive from previous Commander-in-Chiefs. In the last 70 years, we can see presidents favoring more adjectives (e.g., Eisenhower), pronouns (e.g., Clinton), or verbs (e.g., Obama).

When applying an authorship attribution based on the style and the content of the *State of the Union* addresses, the resulting success rate is around 95.7% when some constraints are respected. When the system incorrectly assigns a given speech, the proposed author is often from the same political affiliation and had lived (or lives) in the same time period. When inspecting the reasons justifying an incorrect attribution, we discover interesting justifications. For example, the first Bush speech was delivered before September 11th and therefore presents a content and style different from all the other Bush addresses focusing more on new topics (e.g., terrorism, Iraq, homeland security).

This study can be extended by analyzing the presidential speeches at the lexical level to detect vocabulary and expressions particular to a given author or those common to a few presidents. Based on such information, we can select or generate a brief summary according to different points of view (e.g., for each president, by clusters, according to a timespan, etc.). Moreover, we can examine those speeches according to other perspectives such as based on psychometric measures (Pennebaker, 2011) or according to their anchoring in time and space. Moreover, we can compare these well-prepared speeches with other more spontaneous interventions such as the presidential answers in press conferences. Finally, it could be pertinent to study the differences between these governmental addresses and the electoral speeches at the lexical, thematic, and rhetorical levels.

Acknowledgments

This research was supported, in part, by the Swiss NSF under Grant #200021_149665/1. I would like to thank to thank the anonymous referees for their helpful suggestions and remarks.

References

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

Bartélémy, J.P., & Guénoche, A. (1991). Trees and proximity representations. New York: John Wiley.

Blei, D.M., & Lafferty, J. (2009). Topic models. In A. Srivastava & M. Sahami (Eds.), Topic models, text mining (pp. 71–94). London: Taylor & Francis.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3(1), 993–1022.

Burrows, J.F. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing, 17(3), 267–287.

Calvet, L.-J., & Véronis, J. (2008). Les mots de Nicolas Sarkozy. Paris: Seuil.

Carpenter, R.H., & Seltzer, R.V. (1970). On Nixon's Kennedy style. Speaker and Gavel, 7(2), 41–43.

Francis, W.N., & Kučera, H. (1982). Frequency analysis of English usage: Lexicon and grammar. Boston: Houghton Mifflin.

Hughes, J.M., Foti, N.J., Krakauer, D.C., & Rockmore, D.N. (2012). Quantitative patterns of stylistic influence in the evolution of literature. Proceedings of the PNAS, 109(20), 7682–7686.

Jockers, M.L. (2013). Macroanalysis. digital methods & literary history. Urbana: University of Illinois Press.

Juola, P. (2003). The time course of language change. Computers and the Humanities, 37(1), 77–96.

Juola, P. (2006). Authorship attribution. Foundations and Trends in Information Retrieval, 1(3), 233–334.

Juola, P. (2013). Using the *Google* n-gram corpus to measure cultural complexity. Literary and Linguistic Computing, 28(4), 668–675.

Kaufman, L., & Rousseeuw, P.J. (1990). Finding groups in data: An introduction to cluster analysis. Hoboken: Wiley Interscience.

Kolakowski, M., & Neale, T.H. (2006). The president's state of the union message: Frequently asked questions. Congressional Research Service, 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: Honoré Champion.

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

Lebart, L., Salem, A., & Berry, L. (1998). Exploring textual data. Dordrecht: Kluwer.

Manning, C.D., & Schütze, H. (1999). Foundations of statistical natural language processing. Cambridge: The MIT Press.

Marcus, M.P., Santorini, B., & Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English: The penn treebank. Computational Linguistics, 19(2), 313–330.

Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P., . . . Aiden, E.L. (2011). Quantitative analysis of culture using millions of digitized books. Science, 331(6014), 176–182.

Moretti, F. (2005). Graphs, maps, trees. Abstract models for literary history. London: Verso.

Paradis, E. (2011). Analysis of phylogenetics and evolution with R. New York: Springer.

Pauli, F., & Tuzzi, A. (2009). The end of year addresses of the presidents of the Italian republic (1948–2006): Discourse similarities and differences. Glottometrics, 18, 40–51.

Pennebaker, J.W. (2011). The secret life of pronouns. What our words say about us. New York: Bloomsbury Press.

DOI: 10.1002/asi

Petrocik, J.R., Benoit, W.L., & Hansen, G.J. (2003). Issue ownership and presidential campaigning, 1952–2000. Political Science Quarterly, 118(4), 599–626.
Savoy, J. (2010). Lexical analysis of US political speeches. Journal of Quantitative Linguistics, 17(2), 123–141.
Savoy, J. (2014). Comparative evaluation of term selection functions for authorship attribution. Digital Scholarship in the Humanities, to appear, DOI: dx.doi.org/10.1093/llc/fqt047.
Sebastiani, F. (2002). Machine learning in automatic text categorization. ACM Computing Surveys, 14(1), 1–27.
Shogan, C.J., & Neale, T.H. (2012). The president's State of the Union address: Tradition, function, and policy implications. Congressional Research Service, 7–5700.
Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclid dependency network. Proceedings of HLT-NAACL 2003, 1, 173–180.
Witten, I.A., Frank, E., & Hall, M.A. (2011). Data mining. Practical machine learning tools and techniques. Amsterdam: Morgan Kaufmann.
Zhao, Y., & Zobel, J. (2007). Searching with style: Authorship attribution in classic literature. In G. Dobbie (Ed.), Proceedings ACSC2007 (pp. 59–68). Ballarat: CRPIT.

Appendix

TABLE A.1. List of the US presidents with the number of addresses and their political affiliation.

#	President Name	# Speeches	From	To	Party
1	George Washington	8	1790	1796	Ind.
2	John Adams	4	1797	1800	F
3	Thomas Jefferson	8	1801	1808	D–R
4	James Madison	8	1809	1816	D–R
5	James Monroe	8	1817	1824	D–R
6	John Quincy Adams	4	1825	1828	N–R
7	Andrew Jackson	8	1829	1836	D
8	Martin Van Buren	4	1837	1840	D
9	William H. Harrison	0	1841	1841	Whig
10	John Tyler	4	1841	1844	D
11	James Polk	4	1845	1848	D
12	Zachary Taylor	0	1849	1849	Whig
13	Millard Fillmore	3	1850	1852	Whig
14	Franklin Pierce	4	1853	1856	D
15	James Buchanan	4	1857	1860	D
16	Abraham Lincoln	4	1861	1864	R
17	Andrew Johnson	4	1865	1868	D
18	Ulysses S. Grant	8	1869	1876	R
19	Rutherford B. Hayes	4	1877	1880	R
20	James A. Garfield	0	1881	1881	R
21	Chester A. Arthur	4	1881	1884	R
22	Grover Cleveland	4	1885	1888	D
23	Benjamin Harrison	4	1889	1892	R
24	Grover Cleveland	4	1893	1896	D
25	William McKinley	4	1897	1900	R
26	Theodore Roosevelt	8	1901	1908	R
27	William H. Taft	4	1909	1912	R
28	Woodrow Wilson	8	1913	1920	D
29	Warren Harding	2	1921	1922	R
30	Calvin Coolidge	6	1923	1928	R
31	Herbert Hoover	4	1929	1932	R
32	Franklin D. Roosevelt	12	1934	1945	D
33	Harry S. Truman	7	1947	1953	D
34	Dwight D. Eisenhower	9	1953	1960	R
35	John F. Kennedy	3	1961	1963	D
36	Lyndon B. Johnson	6	1964	1969	D
37	Richard Nixon	5	1970	1974	R
38	Gerald R. Ford	3	1975	1977	R
39	Jimmy Carter	3	1978	1980	D
40	Ronald Reagan	7	1982	1988	R
41	George H.W. Bush	4	1989	1992	R
42	William J. Clinton	8	1993	2000	D
43	George W. Bush	8	2001	2008	R
44	Barack Obama	6	2009	2014	D

When looking at the US political parties, we encounter first the Federalists (F) who will disappear around 1812. G. Washington seats as independent but was close to the Federalists' ideas. It's rival was the Democratic–Republican (D–R) Party that will split in two in 1825 to form the Democrat Party (D) and the National Republican (N–R). This latter party, dissolved in 1833, will be followed by the Whig Party. In 1854, members of the Whig Party founded the Republican (R) Party who takes the lead over the Whig movement.

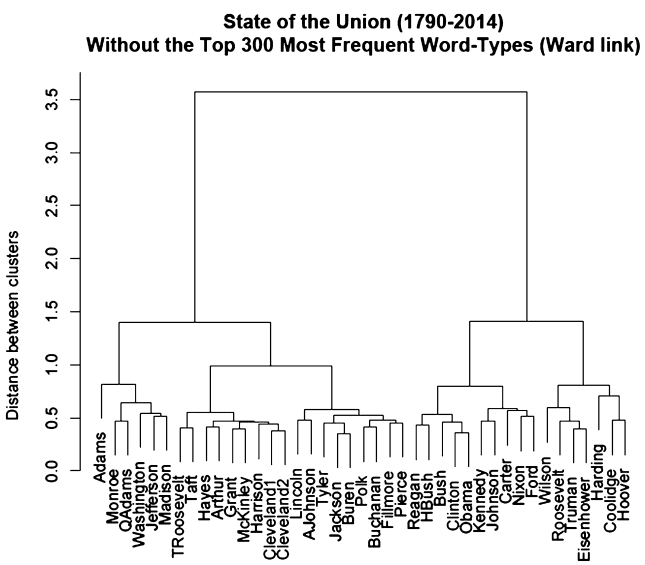

FIG. A.1. Representation using a dendrogram based on the president's profile (Content only, Ward method).