Text Representation Strategies: An Example With the State of the Union Addresses

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

Based on State of the Union addresses from 1790 to 2014 (225 speeches delivered by 42 presidents), this paper describes and evaluates different text representation strategies. To determine the most important words of a given text, the term frequencies (*tf*) or the *tf idf* weighting scheme can be applied. Recently, latent Dirichlet allocation (LDA) has been proposed to define the topics included in a corpus. As another strategy, this study proposes to apply a vocabulary specificity measure (Z score) to determine the most significantly overused word-types or short sequences of them. Our experiments show that the simple term frequency measure is not able to discriminate between specific terms associated with a document or a set of texts. Using the *tf idf* or LDA approach, the selection requires some arbitrary decisions. Based on the term-specific measure (Z score), the term selection has a clear theoretical basis. Moreover, the most significant sentences for each presidency can be determined. As another facet, we can visualize the dynamic evolution of usage of some terms associated with their specificity measures. Finally, this technique can be employed to define the most important lexical leaders introducing terms overused by the *k* following presidencies.

Introduction

When facing a large corpus, we may want to summarize it using a short description or even limit such a synthesis to a few descriptors assigned manually (indexing). Of more interest would be to have an overview that can show the specificities of the different subparts of a text collection. This is the main objective of this study using a political corpus containing 225 State of the Union addresses delivered by 42 U.S. presidents from 1790 (G. Washington) to 2014 (B. Obama).

Received November 17, 2014; revised January 5, 2015; accepted January 6, 2015

 DOI: 10.1002/asi.23510

This corpus shows us the issues and difficulties facing the United States during its existence. Provided on an annual basis, each State of the Union address describes the situation of the country as required by the Constitution. It also indicates the political priorities of the current tenant of the White House and the proposed legislative projects that the Congress should deliberate during the upcoming year.

With this corpus, how can we extract the terms and expressions that can best characterize each president? Can we observe some stylistic elements belonging to a continuous sequence of presidents? Can we detect the topics more closely related to a given presidency or common to a set of presidents or to leaders of the same political party? To answer these questions, we propose to apply a method defining the vocabulary specific to a given subset compared to the whole corpus (Muller, 1992). Based on this approach, we can then define the lexical items and topics specific to a president or a set of presidents. To compare this method with other representation strategies, we will show the differences with extracting schemes based only on the term frequency (*tf*) information, the well-known *tf idf* weighting scheme, or according to the probabilistic latent Dirichlet allocation (LDA) approach.

The rest of this paper is organized as follows. An overview of related work is presented in then next section followed by a section depicting the main features of the corpus used in our experiments. We then describe how we define the vocabulary specific to a given subset and apply it to each presidency. The next section describes and illustrates a selection process able to extract the most specific sentences of a given president. Then we explain how we can visualize the term specificity evolution over time (or across the entire corpus). Finally, the last section shows how we can determine lexical leaders, presidents able to introduce an expression or a formulation that is then overused by their immediate followers.

Related Work

To define the main topics of a given corpus, we can generate an overview using a tag cloud system (e.g., as proposed by the website www.wordle.net). With this approach, a picture of lexical terms is automatically generated from the input text. The font size of each word depends on its occurrence frequency (*tf*) (terms with high frequency are more visible). Moreover, very frequently used words (usually function words such as determiners, prepositions, conjunctions, pronouns, auxiliary verbs, or modal forms) are ignored because they do not usually convey useful meaning. Those words, however, are important to accurately reflect the style of the different authors (Stamatatos, 2009). Finally, the surface form of each word is strictly respected and the system views as distinct the tokens *states*, *state*, and *States*.

Instead of defining the importance of each term according to only its *tf* value, we can consider defining the weight of each word by computing its *tf idf* value, a measure well known in automatic indexing schemes (Manning, Raghavan, & Schütze, 2008), or in text categorization approaches (Sebastiani, 2002). This formulation takes account of the importance of the term inside a document (or a set of texts) by the *tf* component, while the *idf* reflects the term scarcity inside the corpus. This general idea can be implemented in different manners as, for example, *ntf idf* with *ntf* = *tf* / max *tf* in the document (Rajaraman & Ullman, 2012). The experiments done with the State of the Union corpus indicate very similar results between these possible variants.

To automatically extract a synthetic view of a given corpus, we can apply the topic model approach or LDA proposed by Blei, Ng, and Jordan (2003) (Blei & Lafferty, 2009). This approach views the corpus as generated by a probabilistic model. More precisely, each document is considered as containing one or more topics, and thus each document corresponds to a distribution over a set of topics. Each topic is represented as a distribution over all words. Given a corpus and a given number of topics, the LDA system 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 a complete probabilistic description of the underlying corpus. This approach does not, however, provide a way to define directly the vocabulary specific to a given president.

Previously mentioned strategies are usually based on isolated words. When considering phrases, the system can provide better semantics. For example, specifying that *health* is a very frequent word does not provide enough information to clearly define its precise meaning (e.g., *healthcare reform* vs. *preserve the general health*). Thus, selecting the *n*-gram of words might be a solution to achieve a more precise meaning. When extracting pertinent sentences to reflect the particular content of a document, various sentence extraction algorithms have been proposed (Paice, 1990; Nenkova & McKeown, 2011). Certainly, selecting a single sentence provides a better context than a few isolated words. However, the presence of pronominal references (e.g., "*I give it to her*") may hurt its readability. Moreover, when trying to combine several sentences, textual cohesion must be taken into account (e.g., anaphoric references). Such considerations, however, are beyond the main purpose of this article focusing on automatically extracting terms specific to a subset of a large corpus.

Finally, as the target application corresponds to a political corpus, we can mention some related works in this perspective. The works of Labbé and Monière (2003, 2008) present similar objectives by proposing a lexical investigation over a relatively long period of governmental speeches (1945–2000). These studies 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. Similar conclusions were reached for the Italian presidents over the same time period (Pauli & Tuzzi, 2009).

The State of the Union Addresses

The corpus used in our experiments contains 225 speeches delivered by 42 U.S. presidents, from G. Washington (January, 8th, 1790) to B. Obama (January, 28th 2014). The main objective of each address is to inform the Congress and the nation about the state of the country on the one hand and, on the other, to expose the presidential legislative projects for the upcoming year. Thus, we have the same context for each speech across more than 200 years. A more detailed analysis of the form and political functions of these speeches can be found in Shogan and Neale (2012).

Even though the president currently uses more diverse channels to explain his choices (other official remarks, press conferences, interviews, website, and web-mediated communication), the State of the Union address remains the most important annual presidential speech. It is clearly a unique opportunity to present directly to both the Congress and the nation the objectives and projects of the White House (Hoffman & Howard, 2006).

Finally, some of these addresses are well known for explaining an important issue or a political position held for decades such as the Louisiana Purchase (1803), the Monroe Doctrine (1823), the Roosevelt corollary to the Monroe Doctrine (1904), the Four Freedoms (1941), or the War on Poverty (1964). In others, we can find the first occurrence of well-known expressions such as the *axis of evil* (2002).

In this study, we assume that the same author is behind all speeches covering a given presidency, and by extension we assume that the president himself is the author. Of course, this is not exact because we know that behind each well-known politician there is usually a speechwriter. For example, behind Kennedy we find the name of Sorensen (Carpenter & Seltzer, 1970), Favreau behind Obama, and

DOI: 10.1002/asi

even Madison and Hamilton behind some speeches delivered by Washington. But, as 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.

Clearly, even if the president is not the real author, he approves the style and the content. Moreover, to take the latest events into account, the president might change some passages before delivering a message.

This political corpus was generated by downloading all of the speeches from the website www.presidency.ucsb.edu. Two presidents (W.H. Harrison [1841] and J.A. Garfield [1881]) do not have any speeches because their terms were limited to a few months. Cleveland appears twice as president corresponding to his two terms interrupted by B. Harrison's presidency. The Appendix presents, with more details, a complete list of all U.S. presidents with the number of their State of the Union addresses.

Each speech was cleaned up by replacing certain UTF-8 punctuation marks with their corresponding ASCII symbol. When needed, the diacritics found in certain words (e.g., naïve) have been removed and the contracted forms were replaced by their equivalent full forms (e.g., don't into do not).

To represent each address, we can employ the word-tokens (e.g., *choose, chose, chosen* or *markets, market*) or the word-types (lemmas or entries in the dictionary). Using this last form, word-tokens belonging to the same dictionary entry are regrouped under the same word-type (e.g., *choose* or *market* in our previous example). Using this representation for our experiments also has the advantage of ignoring possible variations due to syntax. For example, the two word-types *I* and *me* are not viewed as distinct but are merged under the common headword *I*. A spelling normalization procedure was applied when different forms were present (e.g., *Viet Nam* or *Vietnam*, *al Qaeda*, *al-Qaida* or *Al Qaida*). In such cases, we kept the same spelling to denote the same entity (e.g., *US*, *U.S.*, *U.S.A.*, *United States*).

To define the corresponding word-type to each word-token, the part-of-speech (POS) tagger proposed by Toutanova, Klein, Manning, and Singer (2003) was applied. For each sentence given as input, this system provides the corresponding POS tag to each token. For example, from the sentence "Our energy policy is creating jobs and leading to a cleaner, safer planet." the POS tagger returns "Our/PRP$ energy/NN policy/NN is/VBZ creating/VBG jobs/NNS and/CC leading/VBG to/TO a/DT cleaner/JJR,/, safer/JJR planet/NN ./.." Tags may be attached to nouns (NN, noun, singular, NNS noun, plural), verbs (VB, base form, VBG gerund or present participle, VBZ 3rd-person singular present), adjectives (JJ), comparative adjective (JJR), personal pronouns (PRP), prepositions (IN), determiners (DT), and adverbs (RB). With this information we are then able to derive the word-type by removing the plural form of nouns (e.g., *laws*/NNS → *law*/NN) or by substituting inflectional suffixes of verbs (e.g., *creating*/VBG → *create*/VB). However, when only the plural form is present, we kept this form (e.g., *terrorists*).

After this preprocessing, our U.S. corpus contains 1,964,025 tokens for 20,604 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 2,432 *dis legomena* (word-types occurring exactly twice, representing 11.8% of the vocabulary). The definite determiner *the* (151,814 occurrences) is the most frequent word-type, followed by *of* (98,337), the comma (96,497), *be* (65,705), the full stop (61,777), *to* (60,487), *and* (60,188), and *in* (38,466).

At the speech level, the mean length is 8,731.2 tokens (standard deviation: 5,860). The longest address was written 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 remarks (average of 1,931 tokens per speech) while Taft (1909–1912) is the author, in mean, of the longest addresses (24,655 tokens).

Term Specificity Measure

The writing style of an author can be characterized by the frequency variations of function words or a subset of them (Stamatatos, 2009). Those terms, however, are of limited interest when focusing on the semantic level. On the other hand, each author can also be described by the particular use of some terms or sequences of them. For example, the word *florins* or the expression *British subjects* cannot characterize recent U.S. presidents. But those expressions belong to the specific vocabulary of other presidents (e.g., Washington). Thus, the vocabulary specific to an author can belong to both some functional words (style) and some topical terms. To define this lexical specificity, Muller's method (1992) can be adopted and was used previously as an authorship attribution scheme (Savoy, 2012). In the current study, the target application pursues a larger scale than a few authors in describing the lexical specificities associated with each U.S. president. Moreover, a recent study shows that when applying a clustering algorithm on this corpus, all speeches appearing under the same presidency tend to regroup themselves under the same cluster (Savoy, 2015).

To measure the specificity attached to a term (defined as a word-type or a sequence of word-types in this study), we split the whole corpus into two disjoint portions denoted P_0 and P_1. For a given term t_i, its occurrence frequency in P_0 is denoted tf_{i0}, and in P_1 by tf_{i1}. In the current study, P_0 corresponds to all speeches written during a given presidency, while P_1 denotes all other addresses. Thus, for the entire corpus the occurrence frequency of the term t_i becomes $tf_{i0} + tf_{i1}$. The total number of tokens in part P_0 (or its length) is denoted n_0, similarly with P_1 and n_1, and the length of the entire corpus is defined by $n = n_0 + n_1$.

DOI: 10.1002/asi

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

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

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

For each term, this procedure defines its specificity weight according to the text P_0. Based on the resulting Z score value, we can verify whether this term is used proportionally with roughly the same frequency in both parts (Z score value close to 0). On the other hand, when a term has a positive Z score larger than a fixed threshold δ (e.g., 3), we consider it as *significantly overused* or belonging to the specific vocabulary of P_0. In such a case, the text P_0 contains significantly more occurrences of the corresponding term than expected by a uniform distribution over the whole corpus. A large negative Z score (less than $-\delta$) indicates than the corresponding term is significantly underused in P_0.

When applying this approach, we can consider only isolated words or sequences of such lexical items (e.g., *nuclear weapon* or *healthcare reform*). Although the whole vocabulary can be analyzed, terms appearing only once or a few times do not usually present a noteworthy interest, and thus can be ignored. Moreover, words occurring in a single or a few texts or used by only one or a few authors can be ignored. Of course, it is also possible to add filters to remove other terms (e.g., numbers, punctuation symbols), or words belonging to some part-of-speech categories (e.g., such as function words when the focus is only on topical aspects). Finally, we suggest modeling the term occurrence using a binomial distribution. This viewpoint can be modified by considering a Poisson process or a hyper geometric distribution (Baayen, 2001, 2008).

To have an idea about the most important isolated word-types extracted by different strategies, Table 1 depicts an example based on Reagan's speeches. Words appearing in the first column are the most frequent ones, and as shown in the table, they are related to the style of the corresponding president. Using the *tf idf* formulation (second column), some of the main topics of this presidency appear as, for example, those related to the federal budget (*spending*, *budget*, *deficit*, *program*), the economic major problems (*job*), the main foreign concern (*Soviet*), as well as terms related to the context of these speeches (*tonight*, *America*).

TABLE 1. The 12 most important words according to different weighting schemes with speeches delivered by Reagan (1982–1988).

tf		*tf idf*		LDA		Specific vocabulary	
tf	Word	*tf idf*	Word	Prob.	Word	Z score	Word
1,974	,	88.90	tonight	0.0090	freedom	33.08	we
1,709	the	74.66	spending	0.0068	future	25.80	America
1,626	.	70.47	Soviet	0.0066	work	24.67	spending
1,381	we	64.02	program	0.0065	family	21.51	freedom
1,215	and	59.75	budget	0.0065	budget	21.29	Sandinista
1,068	of	53.77	deficit	0.0063	tonight	20.69	let
1,043	be	52.10	job	0.0060	federal	18.87	Soviet
1,017	to	49.68	dream	0.0059	tax	18.34	deficit
713	an	48.43	percent	0.0059	free	18.24	dream
664	in	48.30	America	0.0058	give	17.49	budget
476	that	47.84	let	0.0056	child	16.67	family
409	for	44.97	help	0.0053	hope	16.21	yes

To provide another example, the Appendix presents a similar table extracted from Obama's speeches (2009–2014).

When applying the LDA method, all functional words (pronouns, articles, prepositions, auxiliary verbal forms, punctuations) have been removed. Those words are very frequent under all presidencies (as shown in the first column of Table 1) and will appear in the higher ranks for all presidents with the LDA approach. Therefore, it will be hard to detect the differences between them. Once those very frequent words are removed, the LDA generative approach was applied and the results for President Reagan are depicted in the third column of Table 1. As other recurrent concerns, we see frequent words related to the family (*family*, *child*), the fiscal policy (*tax*), and recurrent terms associated with general and abstract objectives (*freedom*, *future, hope*). Only two words appear in both the *tf idf* and LDA lists (*tonight*, *budget*).

These three representation schemes do not provide a clear and theoretically grounded decision rule specifying how many terms are significantly associated with a presidency. Different ad hoc rules can be adopted, for example, taking the top k ranking terms or terms having a higher score than a predefined threshold (Rajaraman & Ullman, 2012). As mentioned previously, the LDA approach clearly requires an additional preprocessing to remove functional terms more associated with the style than the content of the speeches.

The specific vocabulary approach has a clear decision rule. We suggest considering overused terms as those having a Z score higher than 3 (and corresponding to 0.14% of the Gaussian distribution). As depicted in the last column of Table 1, the most significant terms associated with Reagan's presidency reflect three aspects: stylistic markers (*we*, *yes*), words associated with the context (*tonight*), and terms related to specific topics of this presidency (*spending*, *Sandinista*, *Soviet*, *deficit*, . . .). It is also worth mentioning than seven words are selected by both the *tf idf* and Z score method over the 12 possible ones (*spending*, *Soviet*, *budget*, *deficit*, *dream*, *America*, and *let*). The intersection with the

DOI: 10.1002/asi

TABLE 2. Five significantly overused terms by 13 different presidents.

Rank	President	Overused terms				
2	Washington	Gentlemen Senate	militia	burthen	Creeks	post office
4	Jefferson	peace friendship	armed vessel	funded debt	Mediterranean	Barbary
5	Jackson	bank united	French	ministry	State bank	Confederacy
9	Polk	Mexico war	Texas	Rio Grande	Paredes	she
1	Lincoln	emancipation	insurgent	slave	rebellion	telegraph
6	T. Roosevelt	man	should	corporation	forest	interstate
7	Wilson	thought	coast submarine	unrest	serviceable	storage
3	Roosevelt	objective	democratic	United Nations	Nazi	nurse
8	Truman	we	atomic	free nation	world	Communist
12	Kennedy	nuclear	Vietnam	recession	common market	balance payment
20	Clinton	we	child	healthcare	21st century	parent
	Bush (son)	Iraq	terrorists	coalition	homeland	weapon
	Obama	job	we	clean energy	why	college

LDA top list is limited to two terms (*freedom* and *family*). These observations indicate that the Z score method produces results closer to the classical *tf idf* paradigm than the LDA approach. Examples coming from Obama's speeches in the Appendix confirm these findings.

Specific Vocabulary of Each Presidency

Instead of describing in detail all of the 42 U.S. presidents that have written a State of the Union address, we will focus on the most important according to Schlesinger's rating (1997). This list, as shown in Table 2, contains the three great presidents (Lincoln, Washington, Roosevelt) and the six near-great (Jefferson, Jackson, T. Roosevelt, Wilson, Truman, and Polk). To have examples from more recent presidents, we have added Kennedy, Clinton, Bush (son), and Obama. To have a more complete picture, we can find in the Appendix a table showing the top five most significant overused word-types under each presidency.

This table shows isolated word-types (e.g., *slave*, *child*, *Vietnam*), as well as bigrams of word-types (e.g., *armed vessel*, *Mexico war*). In the latter case, such sequences are formed by considering their POS tags and by selecting only nouns and adjectives, being adjacent (e.g., *free world*) or separated by function words or punctuation symbol (e.g., *balance (of) payment*). The short context of a word-type may be useful in more precisely determining the meaning of an isolated word such as *nuclear* with a possible association with *weapon* or *power-plant*. Finally, instead of computing the Z score of all possible word-types, we ignore less frequent ones (having an occurrence frequency smaller than 20 in the whole corpus) or those used by only a few presidents (terms appearing in speeches delivered by fewer than four presidents).

First, this table depicts some function words (such as pronouns, auxiliary or modal verbs) associated with a presidency. These words indicate a specific aspect of the presidential style not reused by all other tenants of the White House. For example, and as depicted in Table 2, the pronoun *we* is associated with Truman, Clinton, and Obama, who used it to establish a link with the audience and to try to involve it more. As shown in the Appendix, the pronoun *we* appears as a style-marker for all presidencies after the Second World War. This feature corresponds to a general trend of last presidents towards a more conversational rhetoric promoting an intimacy between the speaker and the audience (Lim, 2002).

T. Roosevelt frequently employs, and in a distinctive way, the verb *should* to invite the Congress to elaborate a new law or to take an action. This aspect is not marginal because T. Roosevelt slowly takes the initiative and leadership over the Congress (Hoffman & Howard, 2006). Finally the pronoun *she* associated with Polk is not related to the feminine gender but refers to Mexico.

The second aspect detected by the specific vocabulary approach is expressions related to the form and context to the State of the Union addresses. For example, Washington starts his speeches with the expression "*Fellow-Citizens of the Senate and House of Representatives*."[1] Different presidents will also reuse this phrase. However, Washington will repeat inside each of his addresses the expressions "*Gentlemen of the Senate*" and "*Gentlemen of the House of Representative*." Therefore this expression appears as specific to the first U.S. President. As another example, we find the term *Tonight* specific to Johnson because he was the first president to utter the State of the Union address in the evening (9 PM Eastern time) in order to achieve a larger television audience.

Third, the most visible aspect of the specific vocabulary detection is the presence of terms related to the topics particular to a given presidency. With Washington (1790–1796), we have the questions related to the *militia* (*for the protection of the frontiers*), the peace with the Indians (*Creeks, Cherokees*), the improvement of the *post office* (and *military post*) across the country, and the concern of additional *burthens* on the community, or the necessity to obtain *loans* (*further loan of 2,500,000 florins has been completed in*

[1]We use italics to indicate terms or phrases appearing in the State of the Union addresses.

DOI: 10.1002/asi

Holland). Such recurrent topics correspond to historical sources about this period (Vincent, 2012).

Under Jefferson's presidency (1801–1808), the White House is confronted with the issue of redemption of the *funded debt* ($8 million, *principal* and *interest*), the need to maintain *peace (and) friendship* with the Indians, and the need to find a solution with the issue of American ships captured by *armed vessels of Spain* and the presence of *Barbary* States in the borders of the *Mediterranean*.

Jackson (1829–1836) has to contend with the renewal of the *Bank of the United States* (which was also one of the main topics during the 1832 election), and the related problems with the *State banks*. Foreign affairs are also usually one of the main competences of the president (e.g., *French ministry*, *French government*). As Jackson is clearly in favor of a strict respect of the rights of the states, the federal level is named sometimes by *Confederacy* or *Confederated States*.

Under Polk's presidency (1845–1848), *Texas* was admitted into the Union. After sending U.S. troops along the *Rio Grande* (the border recognized by Mexico was the Nueces River), the *war* against *Mexico* (*Paredes* was the President of Mexico) was inevitable and turns in favor of the US (with the acquisition of the territories of New Mexico, Arizona, Nevada, Utah, California, Colorado, and Wyoming).

For the other presidents, the terms depicted in Table 2 are relatively obvious. We can, however, mention that Lincoln was in favor of the *Atlantic telegraph* (connecting United States with Europe) as well as its extension in the Pacific region. This communication device was also very effective in winning the Civil War. T. Roosevelt wants to improve *interstate* commerce, and *interstate* business, to regulate and supervise *corporations* (and especially *combination of corporations*) and protect *forest* reserves. With Wilson, we encounter the needs of the war (*coast submarine*) as well as for Roosevelt (*Nazi*, *nurse*).

For the four more recent presidents, the word *nuclear* is associated with force, weapons, and defense. Under Kennedy's presidency, the economic problems appear as top priorities with the terms *recession*, international *balance* (of) *payment*, and the beginning of the *Vietnam* war. With Obama, the word-type *job* (and the unemployment after the 2008 crisis) is clearly the most important priority, with *clean energy* as another concern. The term *healthcare*, however, is first associated with Clinton, who was the first to try to create a universal medical coverage system.

Using the Z score values associated with each term, we can extract the overused terms that are able to characterize a presidency either according to its style (with function words such as *I*, *we*, *why*) or form (e.g., with the introductory and final sentences). More important, this term specificity measure can extract the vocabulary related to the particular issues that the president must face during his term. Clearly, the examples given in Table 2 illustrate the usefulness of the specific vocabulary method when working with many different authors or other subdivisions of a large corpus.

Most Specific Sentences

Based on the previous method, a specificity weight can be associated with each term. Using these weights, a system can compute the specificity of a sentence as the sum of the weights of its components. In our implementation, we suggest to simply sum all significantly overused terms (or terms having a Z score larger than the predefined positive threshold δ, fixed at 3 in this study).

This extraction strategy tends to favor longer sentences over a weighting scheme based, for example, on the average number of overused terms or another mean sentence weight formula. Experiments based on such mean values tend to extract very short sentences (e.g., "*Yes, we can*," "*Thank you*" or "*May God bless America*"). Such short descriptors are not useful to clearly indicate an important theme under a given presidency.

Based on our extraction strategy, the most specific sentence contains many significantly overused word-types or sequences of such word-types. The extracted sentence corresponds usually to one well-known legislative priority or project of the president. Starting with Lincoln, the following sentence was extracted from his 1863 State of the Union address. This sentence is a part of a proposed amendment to the US Constitution.

> The President of the United States shall deliver to every such State bonds of the United States bearing interest at the rate of __ per cent per annum to an amount equal to the aggregate sum of___ for each *slave* shown to have been therein by the Eighth Census of the United States, said bonds to be delivered to such State by installments or in one parcel at the completion of the abolishment, accordingly as the same shall have been gradual or at one time within such State; and interest shall begin to run upon any such bond only from the proper time of its delivery as aforesaid.

In this example, the significantly overused terms by Lincoln are underlined and italics are used to denote terms appearing in the top 10 of the most overused terms (e.g., *slave*).

With Roosevelt, the most specific sentence appears in 1945 and summarizes the four freedoms.

> Our own *objectives* are clear; the *objective* of smashing the militarism imposed by *war* lords upon their enslaved peoples the *objective* of liberating the subjugated Nations, the *objective* of establishing and securing freedom of speech, freedom of religion, freedom from want, and freedom from fear everywhere in the world.

The most characteristic sentence of Kennedy's speeches is included in the 1962 State of the Union address. This passage shows the importance of economic considerations under this presidency. It clearly indicates that the president is also playing the role of the legislative leader. Finally, Kennedy's style (Carpenter & Seltzer, 1970) can be typified by long sentences and this example also illustrates this aspect.

DOI: 10.1002/asi

> To expand *our* growth and job opportunities, I urge on the Congress three measures: First, the Manpower Training and Development Act, to stop the waste of able-bodied men and women who want to work, but whose only skill has been replaced by a machine, or moved with a mill, or shut down with a mine; Second, the Youth Employment Opportunities Act, to help train and place not only the one million young Americans who are both out of school and out of work, but the twenty-six million young Americans entering the labor market in this decade; and Third, the 8 percent tax credit for investment in machinery and equipment, which, combined with planned revisions of depreciation allowances, will spur *our* modernization, *our* growth, and *our* ability to compete abroad.

Reagan's speeches can be characterized by an overuse of pronouns *we*/*us*/*our*, the verb *do*, the function words *what*, *here*, *just*, *but*, and the use of the genitive case in the form of the *'s*. Of course, we can also find more topical terms such as *child, America*, *freedom*, or *deficit* (as shown in Table 1). The system extracts the following most specific sentence from the 1983 State of the Union address.

> 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.

With Obama, the economic and financial themes possess a central place and the related terms such as *job*, *business*, and *tax* are significantly overused under his presidency. The most specific sentence extracted from the speech delivered in 2013 is the following:

> The *American* people deserve a tax code that helps small businesses spend less time filling out complicated forms, and more time expanding and hiring; a tax code that ensures billionaires with high-powered accountants can not pay a lower rate than their hard-working secretaries; a tax code that lowers incentives to move *jobs* overseas, and lowers tax rates for businesses and manufacturers that create *jobs* right here in America.

These examples illustrate the usefulness of detecting specific sentences according to a given author. Each of them presents a mix between terms more related to the style (such as pronouns, modal verbs, or other function words) and word-types reflecting one or more of the characteristic concerns of a given presidency. Moreover, favoring long sentences tends to reduce the presence of anaphoric references across sentences and thus increases the readability of the selected sentences.

Finally, instead of being limited to a single sentence, the system may return a few specific sentences related to a given presidency. To achieve this an iterative process can be applied. To define the most specific sentence, we follow the previously described extraction scheme. Then for all terms appearing in this sentence their specific values are reduced by a fixed amount (e.g., six in this study). This decrease tends to promote other overused terms reflecting another recurrent topic. We then iterate with the sentence extraction scheme to determine the next most specific sentence. When applying this procedure to Reagan's speeches, Table 3 depicts the five most specific sentences. In this table, the first column indicates the year of the State of the Union address. A similar example with Obama's speeches is depicted in the Appendix.

TABLE 3. The five most significant sentences from the State of the Union addresses uttered by Reagan (1982–1988).

Year	Specific sentences
1983	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.
1988	And as we have worked together to bring down spending, tax rates, and inflation, employment has climbed to record heights; America has created more jobs and better, higher paying jobs; family income has risen for 4 straight years, and America's poor climbed out of poverty at the fastest rate in more than 10 years.
1982	And then there are countless, quiet, everyday heroes of American who sacrifice long and hard so their children will know a better life than they've known; church and civic volunteers who help to feed, clothe, nurse, and teach the needy; millions who've made our nation and our nation's destiny so very special, unsung heroes who may not have realized their own dreams themselves but then who reinvest those dreams in their children.
1986	Tonight the American people deserve our thanks for 37 straight months of economic growth, for sunrise firms and modernized industries creating 9 million new jobs in 3 years, interest rates cut in half, inflation falling over from 12 percent in 1980 to under 4 today, and a mighty river of good works, a record $74 billion in voluntary giving just last year alone.
1984	Can we love America and not reach out to tell them: You are not forgotten; we will not rest until each of you can reach as high as your God-given talents will take you.

Dynamic Evolution of Some Specific Terms

The previous sections demonstrate the usefulness of the application of the specific vocabulary method to describe distinctive aspects of a given presidency or to detect their most specific sentences. This term specificity measure can be used to visualize the dynamic evolution of selected terms over time or across the different presidencies.

Figure 1 shows the Z score evolutions of the word-types *job*, *tax*, *debt*, and *bank*. In this figure, the two horizontal dashed lines represent the limits (±3) between which the variations must be interpreted as normal fluctuations.

The term *job* is depicted with a dashed black line in Figure 1. This term appears usually below the second horizontal limit (indicating the limit value of −3) and thus corresponds to a significantly underused term. From

DOI: 10.1002/asi

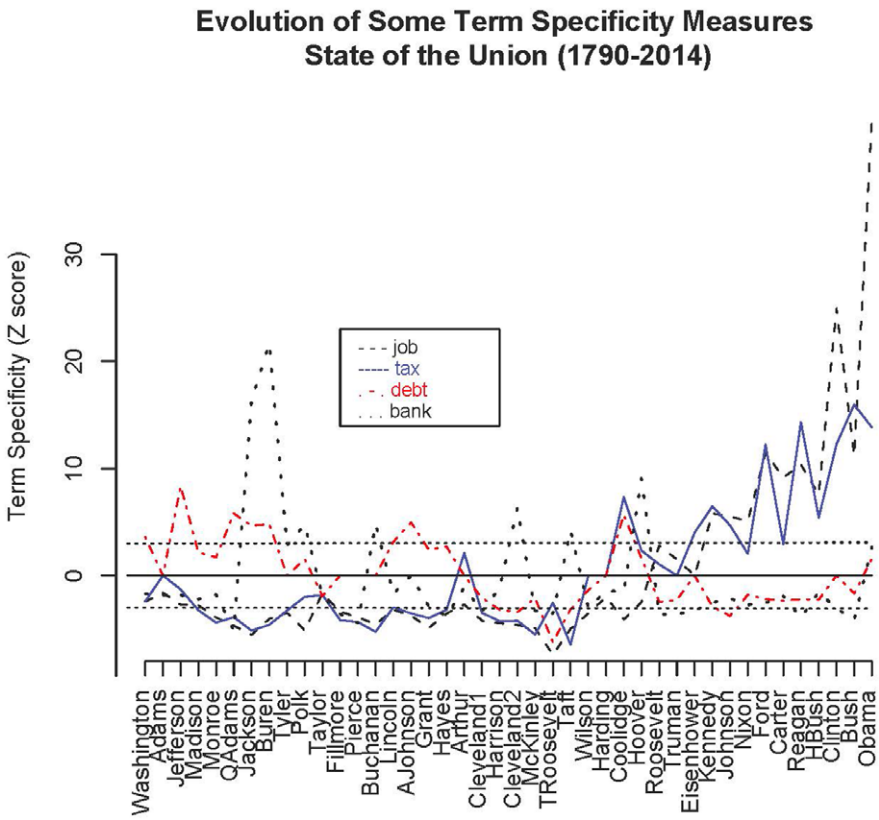

FIG. 1. Variations of the term specificity measure (Z score) of some terms across the State of the Union corpus (1790–2014). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Eisenhower's presidency, this word-type was significantly overused by all presidents, with a first maximum reached under Clinton and the highest value under Obama's presidency.

The term *tax* shown with a thin solid blue line in Figure 1 follows mainly the same pattern. Coolidge (1923–1928) was the first to significantly overuse this word-type. Starting with Kennedy, the evolution of this term follows a sinusoidal curve, with maximums reached under Ford, Reagan, and Bush (son) (and his policy of *tax* reduction).

The evolution of the usage of the term *debt* is displayed with a dotdashed red line. This word-type is overused by different presidents during the first 50 years of existence of the United States. Jefferson (1801–1808), then Quincy Adams (1825–1828), Jackson (1829–1836), and Van Buren (1837–1840) were faced with this issue. Of course, this question reappears under A. Johnson's presidency (1865–1868) just after the Civil War (that had to be funded). Coolidge (1923–1928) is the last president who overused this term. This word-type is not absent in current speeches but this issue appears more under the term *deficit*.

The word-type *bank* follows another pattern. As shown in Figure 1 with a dotted black line, this term is clearly overused by Jackson (1829–1836) but the highest value is achieved under Van Buren's presidency (1837–1840) (with the issue of the *Bank of the United States*, currently the *Federal Reserve*, established under Wilson's presidency). This word-type also appears overused during Buchanan's term (1857–1860), the second term of Cleveland (1893–1896), Taft's term (1909–1912), and Hoover's presidency (1929–1932) (with issues related to some bank bankruptcies during the Great Depression).

If we consider the longest sequence of overused terms, we can find the word-type *need* and the full stop, two terms systematically overused by all 15 presidents after Coolidge (1923). From this presidency, the sentences tend to be shorter and thus we require more full stops. This modification can also be explained by the fact that, from 1934, the presidents have usually delivered this address in spoken form.

The word-types *we*, *can*, and *world* are overused by all the 13 presidents starting from Roosevelt (1933). Except *world*, these terms are mainly related to the style of the president. The auxiliary verb *should*, overused by T. Roosevelt (1901–1908), tends to be replaced by *need* and *can*. This is an indication that the presidents of the second half of the 20th century tend to have a stronger position to face Congress than their predecessors of the second half of the 19th century.

With the frequent use of the pronoun *we*, the president tends to establish a link with the audience. In Table 2, this pronoun appears three times and in association with Truman, Clinton, and Obama. Of course, the pronoun *we* can refer to different entities such as *I and my government*,

DOI: 10.1002/asi

I and you (the Congress), *our country*, or simply together. Another explanation of the more frequent usage of this pronoun is the fact that the State of the Union addresses are, from 1947, televised and, from 1965, they are delivered in the evening. Clearly, the president wants to reach a larger audience, and addresses himself more directly to the people of America to find external support for his proposals. Such a strategy might be important when the president does not have a majority in the Congress.

The pronoun *I* was overused by all nine presidents from Johnson on (1964). Clearly, the speeches tend to be less distant, warmer, honest, and personal (Pennebaker, 2011). The president also clearly indicates that *he* is the leader. The pronoun *you* is also overused but only for the last five presidents, starting with Reagan (1982).

When the president needs to designate his country, he may use different forms such as *U.S.*, *United States*, our *country*, the *nation*, the *Union*, or simply *America*. This last form, together with *American*, are significantly overused by all the presidents from Johnson on (1964). The previous presidents tend to opt more for other formulations such as *nation* with Roosevelt, Truman, or Kennedy, *Union* for Lincoln (1861–1864), or *United States* for Washington (1790–1796), Monroe (1817–1824), or Grant (1869–1876). In France, the preferred term is *République* but the words *nation* or simply *France* are also frequently used (Labbé & Monière, 2003) while Italian presidents tend to use *Italia* more often than *Repubblica* (Pauli & Tuzzi, 2009).

When we are looking backward, the definite determiner *the* is significantly overused by 11 consecutive presidents, from Madison (1809) to Buchanan (1860) and then by 10 tenants of the White House from A. Johnson (1865) to Taft (1912). Only Lincoln (1861–1865) does not overuse this determiner. The political style of the 19th century is clearly associated with more nouns typically used to describe and explain the situation.

Inspecting more of the content of those 19th century speeches, we discover sequences of overused word-types such as *bond* (with the issue of the debt management, from Hayes [1877] to McKinley [1900]), *Indian* and *reservation* (from Grant [1869] to Cleveland [1896]), *silver* (in the sense of *silver coin*, from Hayes [1877] to Cleveland [1896]), or *commerce* (from Adams [1797] to Jackson [1836]).

When analyzing some recent sequences of significantly overused terms, a picture of recurrent issues appears. Unemployment is certainly one of these problems. The word-type *job* is overused by all presidents from Kennedy on (1960) while the terms *work* and *worker* are overused from Carter on (1978). *Spending* is significantly overused from Nixon on (1970) and can be associated with other series of overused terms such as *cut* (from Ford [1975]), *budget* (from Eisenhower [1953]), and more recently with *tax* (from Reagan [1982]). As sequences of overused terms, there are other persistent topics such as *family*, *parent*, *child*, *senior*, *woman*, and *school*. As a more ambiguous word-type, we have the term *nuclear* significantly overused from Ford (1975) and usually associated with weapons (arms, arsenal, proliferation) and currently it is more linked with power-plants. As additional sequences of significantly overused word-types, we can observe *reform*, *stop*, *strategy*, *research*, *technology*, *leader*, and *commitment*.

Lexical Leaders

Using the term specificity measure, we can also determine the presidents who present or explain a new issue using an overused term that will then be overused by the k following presidencies. Based on this definition, the president who is the first to overuse this term is called the lexical leader. Such an expression may occur due to a new issue or problem (e.g., *inflation*, *Al Qaida*), recurrent over at least one presidency (with $k = 1$) or more (e.g., with $k = 3$, covering roughly two decades). Of course, fixing a large value for k will decrease the probability of observing such an overused term sequence.

In this study the value of k was fixed to 3, which implies that the sequence covers at least four presidents overusing the same expression or word-type. For example, as a figure of style, Reagan overused the term *bless* and *God*, both will then be overused by the following four presidents (until Obama). For example, Reagan or Bush (father) repeat it many times and may finish their speeches with the phrase "*God bless you, and God bless America*" while Clinton or Obama tend to only utter the formulation "*God bless you, and God bless the United States*" once. Such an expression is not fully absent from the other president's writings, but it is less frequent and thus not overused. For example, we can mention Truman with the final phrase "*May God bless our country and our cause*" or Roosevelt with "*God must forever bless*." In the beginning of the 19th century, the reference to *God* appears differently, as for example, with Jackson (1829) who wrote "*I now commend you, fellow citizens, to the guidance of Almighty God*." But with Reagan and his followers, the term *God* appears in other places in their remarks and the last sentence in Table 3 is such an example.

To speed up the computation, we added the constraint that the word-type must have an occurrence frequency of 20 or more over all the State of the Union addresses. With this restriction, the vocabulary contains 4,155 distinct word-types. Fixing the value $k = 3$, we can detect 160 word-types that are overused under four consecutive presidencies. The longest sequence is with the full stop, starting with Coolidge (1923) until Obama (2014). This aspect is clearly related to a rhetoric evolution towards the use of shorter sentences and a style more direct, without long explanations (Lim, 2002). With $k = 3$, Bush (father) is the last president that can be considered as a lexical leader (Clinton is only followed by two presidents, Bush [son] and Obama).

Inspecting the distribution of these overused term sequences, we can count 31 such series from Washington (1790) to Hoover (1932), and 129 from Roosevelt (1933) to Obama (2014). This skewed distribution indicates that we

DOI: 10.1002/asi

can distinguish between two main periods, before and after 1933.

As lexical leaders until 1933, the three most influential presidents are Hayes (with five terms, namely, *bond*, *silver*, *coinage*, *sinking*, and *reservation*), Coolidge (1923–1928) (five word-types: *economic*, *national*, *through*, *need*, and the full stop), Grant (1869–1876) (four terms: *polygamy*, *Indian*, *subject*, and *commend*), and Arthur (1881–1884) (four word-types: *suggest*, *suggestion*, *revenue*, and *pension*). The other presidents have introduced between 0 to two terms overused by the next three presidencies, such as Adams with the word-types *communication* and *commerce*. Until Roosevelt (1933), the different tenants of the White House preferred to use their own formulations. On the other hand, when they are using the same terms as their predecessors, those forms are usually not significantly overused. As another possible explanation, we must recall that those speeches were mainly only written.

From 1933, we observe a greater number of sequences of overused terms. The most fruitful lexical leader is Reagan (with 52 word-types as, for example, *school*, *initiative*, *tax*, *future*, *tell*, *say*, *technology*, *woman*, *reform*, etc.), followed by Bush (father) (with 17 word-types, e.g., *student*, *lead*, *clean*, *health*, *care*, *teacher*, *Israel*, *need*, etc.), Roosevelt (10 word-types: *world*, *peace*, *defense*, *goal*, *task*, *program*, *can*, *must*, *we*, and *today*), and Truman (with 10 word-types: *Soviet*, *communist*, *progress*, *major*, *help*, *move*, *increase*, *level*, *billion* and *basic*).

Following this group, we can find Johnson (seven terms: *people*, *America*, *American*, *challenge*, *will*, *I*, and *face*), Ford (seven word-types: *must*, *strategic*, *cut*, *down*, *economic*, *nuclear*, and *Lincoln*), Carter (seven terms: *work*, *worker*, *build*, *together*, *leader*, *commitment*, *young*, *tonight*), Nixon (six word-types: *strong*, *inflation*, *spending*, *growth*, *decade*, and *problem*), Eisenhower (five terms: *job*, *budget*, *new*, *space*, and *quest*), and finally Kennedy (three terms: *more*, *job*, *percent*). From a lexical point of view, the presidencies of Eisenhower, Kennedy, or Ford present only a few overused terms reused by the following presidents. When analyzing the style of the U.S. presidents (Savoy, 2015), we can see that these three presidents are strongly related to only one other president and relatively distant from the others. In other words, and from a lexical point of view, they are isolated. On the other hand, Reagan's style is strongly associated with Clinton, Obama, and the two Bush presidents.

Conclusion

The State of the Union corpus contains the annual speeches of 42 U.S. presidents over more than 200 years. Each address depicts the situation of the country on an annual basis and presents the legislative agenda and priorities of the White House for the forthcoming year. This corpus provides a pertinent collection to inspect the terms or expressions frequently used over time or the main political formulations and topics according to each president.

Based on this corpus, we have explained how we can measure the term specificity according to a given president. This measure is then able to detect terms (isolated words or short sequence of *n*-gram of words) particular to a given presidency or to a given time period. When compared to other representation strategies, the specific vocabulary scheme proposes a clear decision rule to determine which terms are overused. This selection strategy shares some similarities with the *tf idf* weighting scheme but both produce different results than either the simple *tf* weighting scheme or the LDA model.

Having associated a specificity weight to each term, the most distinctive sentence of each president can be determined. Examples show us that such sentences tend to reflect both the president's style and one of his main concerns. If needed, the suggested extraction scheme may produce a few significant sentences providing a better overview of the most recurrent concerns of a given presidency.

As another facet, the dynamic evolution of terms can be analyzed. For example, the frequency of the definite article *the* and the preposition *of* tends to decrease over time, while the use of the full stop tends to increase. The presidential sentences tend to be shorter, with fewer nouns and therefore present less complex explanations. From Roosevelt (1933), the frequencies of personal pronouns (*we*, *I*, *you*) tend to increase significantly. When inspecting more topical terms, we can detect different patterns. For example, the word-type *debt* was significantly overused during the first 50 years and then it appears less in subsequent governmental speeches. As an opposite example, the words *job* or *tax* are significantly overused by the recent presidents.

Finally, we suggest defining the most prolific presidents by considering terms significantly overused during a given period of *k* presidencies. According to this lexical ranking, Reagan appears in the first position, followed by Bush (father) then Roosevelt and Truman. Our analysis indicates, however, that presidents coming after 1933 are inclined to overuse significantly more terms used by their predecessors. This finding suggests that the last presidencies tend to have more similar speeches than presidents of the 19th or beginning of the 20th century.

Acknowledgments

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

References

Baayen, H.R. (2001). Word frequency distribution. Dordrecht, Netherlands: Kluwer.

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

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. The Journal of Machine Learning Research: JMLR, 3(1), 993–1022.
Carpenter, R.H., & Seltzer, R.V. (1970). On Nixon's Kennedy style. Speaker and Gavel, 7(2), 41–43.
Hoffman, D.R., & Howard, A.D. (2006). Addressing the State of the Union. The evolution and impact of the president's big speech. Boulder, CO: Lynne Rienner.
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, Canada: Monière-Wollank.
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., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge, UK: Cambridge University Press.
Muller, C. (1992). Principes et méthodes de statistique lexicale. Paris: Honoré Champion.
Nenkova, A., & McKeown, K. (2011). Automatic summarization. Foundations and Trends in Information Retrieval, 5(2–3), 1–134.
Paice, C.D. (1990). Constructing literature abstracts by computer: Techniques and prospects. Information Processing & Management, 26(1), 171–186.
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(1), 40–51.
Pennebaker, J.W. (2011). The secret life of pronouns. What our words say about us. New York: Bloomsbury Press.
Rajaraman, A., & Ullman, J.D. (2012). Mining of massive datasets. Cambridge, UK: Cambridge University Press.
Savoy, J. (2012). Authorship attribution based on specific vocabulary. ACM Transactions on Information Systems, 30(2), 170–199.
Savoy, J. (2015). Text clustering: An application with the State of the Union addresses. Journal of the American Society for Information Science and Technology: JASIST, DOI: 10.1002/asi.23283.
Schlesinger, A.M., Jr. (1997). Rating the US presidents: Washington to Clinton. Political Science Quarterly, 11(2), 179–190.
Sebastiani, F. (2002). Machine learning in automatic text categorization. ACM Computing Survey, 34(1), 1–27.
Shogan, C.J., & Neale, T.H. (2012). The president's State of the Union address: Tradition, function, and policy implications (pp. 7–5700). Washington (DC): Congressional Research Service.
Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology: JASIST, 60(3), 538–556.
Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclid dependency network. In M. Hearst & M. Ostendorf (Ed.), Proceedings of NAACL 2003 (pp. 252–259). Edmonton, Canada: ACL.
Vincent, B. (2012). Histoire des Etats-Unis. Paris: Flammarion.

Appendix

TABLE A1. List of the U.S. presidents with the number of the State of the Union addresses (1790–2014).

#	President name	# speeches	From	To
1	George Washington	8	1790	1796
2	John Adams	4	1797	1800
3	Thomas Jefferson	8	1801	1808
4	James Madison	8	1809	1816
5	James Monroe	8	1817	1824
6	John Quincy Adams	4	1825	1828
7	Andrew Jackson	8	1829	1836
8	Martin Van Buren	4	1837	1840
9	William H. Harrison	0	1841	1841
10	John Tyler	4	1841	1844
11	James Polk	4	1845	1848
12	Zachary Taylor	1	1849	1849
13	Millard Fillmore	3	1850	1852
14	Franklin Pierce	4	1853	1856
15	James Buchanan	4	1857	1860
16	Abraham Lincoln	4	1861	1864
17	Andrew Johnson	4	1865	1868
18	Ulysses S. Grant	8	1869	1876
19	Rutherford B. Hayes	4	1877	1880
20	James A. Garfield	0	1881	1881
21	Chester A. Arthur	4	1881	1884
22	Grover Cleveland	4	1885	1888
23	Benjamin Harrison	4	1889	1892
24	Grover Cleveland	4	1893	1896
25	William McKinley	4	1897	1900
26	Theodore Roosevelt	8	1901	1908
27	William H. Taft	4	1909	1912
28	Woodrow Wilson	8	1913	1920
29	Warren Harding	2	1921	1922
30	Calvin Coolidge	6	1923	1928
31	Herbert Hoover	4	1929	1932
32	Franklin D. Roosevelt	12	1933	1945
33	Harry S. Truman	7	1947	1953
34	Dwight D. Eisenhower	9	1953	1960
35	John F. Kennedy	3	1961	1963
36	Lyndon B. Johnson	6	1964	1969
37	Richard Nixon	5	1970	1974
38	Gerald R. Ford	3	1975	1977
39	Jimmy Carter	3	1978	1980
40	Ronald Reagan	7	1982	1988
41	George H.W. Bush	4	1989	1992
42	William J. Clinton	8	1993	2000
43	George W. Bush	8	2001	2008
44	Barack Obama	6	2009	2014

Table A2 shows the top 12 word-types selected by the four strategies based on Obama's speeches. As for the previous example with Reagan's addresses (Table 1), the simple term frequency (*tf*) does not provide pertinent information. The *tf idf* weighting scheme indicates more clearly some important topics related to Obama's presidency with lemmas such as *job*, *kid*, *college*, *student*, *innovation*, and *deficit*. Other words are related to the form (*tonight*) or the style (*why*, *get*, *let*). The last column (Z score) forms a list related to the *tf idf* one (8 terms in common over 12). As distinct terms, we can found some stylistic items (*we*, *do*) and the topical terms *energy* and *Afghan*. It is interesting to observe that the pronoun *we* appears also as the fourth most

TABLE A2. The top most important words according to different weighting schemes with speeches uttered by Obama (2009–2014).

tf		*tf idf*		LDA		Specific vocabulary	
tf	Word	*tf idf*	Word	Prob.	Word	Z score	Word
2291	,	219.24	job	0.0129	work	46.68	job
2216	.	131.76	tonight	0.0128	job	35.00	we
1823	the	98.63	get	0.0093	tax	30.61	why
1695	we	81.88	help	0.0087	business	30.06	get
1406	be	78.53	kid	0.0080	cut	24.43	college
1383	and	77.27	college	0.0065	plan	23.03	kid
1360	to	72.42	why	0.0064	family	22.65	energy
1011	of	65.99	student	0.0061	health	21.34	innovation
915	an	56.87	deficit	0.0060	give	21.15	student
887	that	56.33	clean	0.0060	economy	21.10	tonight
648	in	54.73	innovation	0.0057	change	20.95	do
530	have	51.52	let	0.0053	care	20.92	Afghan

TABLE A3. The five most significant sentences from the State of the Union addresses uttered by Obama (2009–2014).

Year	Specific sentences
2013	The American people deserve a tax code that helps small businesses spend less time filling out complicated forms, and more time expanding and hiring; a tax code that ensures billionaires with high-powered accountants can not pay a lower rate than their hard-working secretaries; a tax code that lowers incentives to move jobs overseas, and lowers tax rates for businesses and manufacturers that create jobs right here in America.
2014	And when our children's children look us in the eye and ask if we did all we could to leave them a safer, more stable world, with new sources of energy, I want us to be able to say yes, we did.
2009	Now, there will be many different opinions and ideas about how to achieve reform, and that is why I am bringing together businesses and workers, doctors and health care providers, Democrats and Republicans to begin work on this issue next week.
2012	Tonight, my message to business leaders is simple: Ask yourselves what you can do to bring jobs back to your country, and your country will do everything we can to help you succeed.
2012	The new rules we passed restore what should be any financial system's core purpose: Getting funding to entrepreneurs with the best ideas, and getting loans to responsible families who want to buy a home, start a business, or send a kid to college.

frequent word in Obama's speeches as depicted in the first column. After removing very frequent words, the LDA list proposes other main concerns of this presidency, namely *tax*, *business*, *health*, *economy* as well as ambiguous word-types (*cut*, *plan*).

Table 3 gives the five most specific sentences of Reagan's speeches. To establish a parallel, Table A3 illustrates the five most specific sentences of Obama's addresses. The first is clearly related to *tax*, *business*, and *jobs* while the second is related to the *clean energy* issue. The third sentence is related the *healthcare reform*. The fourth indicates that one of most important issues under this presidency is related to *new jobs*, and the last one concerns the *financial* system that must serve both *business* and *families*.

DOI: 10.1002/asi

TABLE A4. The top-five most significantly overused isolated words under each presidency.

Presidency	Word 1	Word 2	Word 3	Word 4	Word 5
Washington	Gentlemen	militia	you	Pennsylvania	burthen
Adams	Gentlemen	Philadelphia	commissioner	amity	capture
Jefferson	funded	Mediterranean	millions	Orleans	Barbary
Madison	British	enemy	militia	savage	council
Monroe	Spain	likewise	presume	colony	adventurer
QAdams	of	enumeration	the	discriminating	Parliament
Jackson	French	bank	Chamber	France	the
Buren	bank	money	suspension	public	institution
Tyler	Texas	she	destined	paper	annexation
Polk	Mexico	Texas	Mexican	she	Paredes
Taylor	California	empire	construct	Granada	observance
Fillmore	California	expedition	postage	duty	dock
Pierce	state	of	compact	territory	sectional
Buchanan	Kansas	constitution	slavery	whilst	slave
Lincoln	emancipation	insurgent	slave	rebellion	telegraph
AJohnson	constitution	rebellion	paper	depreciated	inclusive
Grant	expatriation	herewith	etc.	of	claim
Hayes	coinage	the	silver	Indian	tender
Arthur	merchandise	likely	receipts	steel	lately
Cleveland1	reservation	silver	coinage	coined	of
Harrison	elector	meat	silver	steamship	$
Cleveland2	gold	note	$	silver	inch
McKinley	island	Cuba	Manila	Puerto	the
TRoosevelt	man	should	corporation	forest	interstate
Taft	canal	wool	court	the	department
Wilson	thought	unrest	play	submarine	storage
Harding	readjustment	railway	manager	transportation	relationship
Coolidge	marketing	agriculture	consolidation	ought	league
Hoover	depression	construction	federal	agency	unemployment
Roosevelt	war	objective	Japanese	fight	democracy
Truman	we	atomic	world	Communist	Soviet
Eisenhower	program	economic	we	Communist	federal
Kennedy	alliance	we	Communist	nuclear	recession
Johnson	Vietnam	we	billion	tonight	poverty
Nixon	America	goal	we	responsive	truly
Ford	energy	oil	program	I	federal
Carter	Salt	inflation	Soviet	we	nuclear
Reagan	we	America	spending	freedom	let
GHWBush	we	tonight	America	I	kid
Clinton	we	child	parent	21st	you
GWBush	Iraq	terrorists	we	coalition	homeland
Obama	job	we	why	get	energy

DOI: 10.1002/asi