Yes, Ronald Reagan's Rhetoric Was Unique—But Statistically, *How* Unique?

CHERYL SCHONHARDT-BAILEY
London School of Economics and Political Science

EDWARD YAGER
Western Kentucky University

SAADI LAHLOU
London School of Economics and Political Science

We use automated textual analysis to compare Ronald Reagan's rhetoric with that of presidents Woodrow Wilson through Barack Obama, using their State of the Union speeches. We are able to assign statistical significance to the thematic content, and to depict spatially the shifting dimensionality in themes used by presidents. We find strong evidence for Reagan's usage of the civil religion rhetoric: over half (59%) of the discourse in his seminal and 48% in his State of the Union speeches focus on civil religion. We also find an apparent shift in modern presidential rhetoric, from themes concerned with (1) institutions, to ones focused more on (2) individuals, families, and children.

While religious affiliation has long been considered to be "one of the most accurate, and least appreciated" indicators of political affiliation (Gallup and Castelli 1989, 249), political scientists paid scant attention to religion in U.S. politics until recent years, ascribing it "second-class status" as an area of study (Wolfe 2010, 20). However, as religion is coming to shape more features of the American political landscape, attention is turning toward measuring its significance. For instance, some have interpreted its growing importance as part of a larger phenomenon of increasing partisan polarization,

with highly religious Americans concentrated at the Republican right of the political spectrum and secular Americans at the Democratic left. According to this view, Christian conservatives in the Republican Party have shifted the party to the right on social issues, while the Democratic Party is moving to the left on social issues, to reflect its relatively larger contingent of secular members. The end product is a political party division between religious and secular Americans, with a shrinking pool of moderately religious Americans (Putnam and Campbell 2010; Wilcox and Robinson 2007).

Yet prior to this phenomenon of religious polarization, religion had not been entirely absent from the literature on American politics: from the 1960s onwards, the concept of *civil religion* has provided insight into and discussion on the role of religion in American politics. While the concept had origins in Jean-Jacques Rousseau's *Social Contract*, Robert Bellah (1967) revived its usage, using evidence particularly from presidential inaugural addresses (Bellah 1967). As heads of state, American presidents are seen as the prophets or ceremonial priests of civil religion, especially as they invoke divine assistance and guidance (Chidester 1988, 91; Fowler, Hertzke, and Olson 1999, 116). Bellah (1967) argued that American civil religion is nondenominational, and so when presidents use *God* and other religious phrases and imagery, they are not conveying their own private religion, but rather appealing to a religious common denominator shared by the majority of Americans. Clearly this supposed religious common denominator is undergoing a sea-change if we are in the midst of the more recent phenomenon of political/religious polarization. If America is dividing into highly religious Republicans and secular Democrats, where is the religious common denominator? In the midst of this great schism sits the U.S. president, whose language and rhetoric has become more closely scrutinized for its religious content and meaning.

A second—and much larger—literature on the presidency of Ronald Reagan has drawn upon the concept of civil religion to help explain Reagan's reputation as the "Great Communicator." While some have attributed this reputation to Reagan's acting experience (Auer 1992), his use of figurative rhetoric (Jasinski 1992), or key "rhetorical moments" that transformed him from actor to president (Blankenship and Muir 1992), a common theme is that Reagan was distinctive among recent American presidents in his use of civil religion rhetoric—and this is key to understanding his acquired reputation as the Great Communicator (Ritter and Henry 1992, 121). He is said to have turned civil religion rhetoric "into a formidable political weapon. In using it he rendered mute those who would oppose him" (Weiler and Pearce 1992, 29).

Within the civil religion literature on U.S. presidents—and particularly those explanations for Reagan's Great Communicator reputation that point to his use of civil religion rhetoric—are three problems of measurement: (1) the empirical evidence is usually anecdotal rather than systematic; (2) without systematic evidence, it is unclear just how unique Reagan's rhetoric (and his apparent reliance on civil religion) was from other recent presidents; and (3) to the extent that evidence of civil religion is found in speeches of U.S. presidents, it is usually from inaugural addresses, which by their very nature are designed to set out the new president's broader principles and "presidential commitment to the country's basic principles" (Campbell and Jamieson

1985) and so are biased in favor of finding civil religion rhetoric. We address these three problems by employing automated textual analysis software to measure statistically the civil religion rhetoric in Reagan's seminal speeches, his State of the Union speeches alone, and then the State of the Union speeches of all modern presidents from Woodrow Wilson to Barack Obama. By using State of the Union speeches as our basis of comparison for Reagan's use of civil religion, we are employing a more difficult challenge for finding evidence of this rhetoric. Unlike inaugurals, State of the Union speeches are constitutionally mandated and, as such, they usually identify what the incumbent president views as problems and then propose legislative action and other means to address these problems (Campbell and Jamieson 1985). Hence, unlike inaugurals, State of the Union speeches tend to be more policy prescriptive and so are less likely to embellish upon the broader, more principled rhetoric of civil religion. (We thus follow Campbell and Jamieson's assessment that State of the Union addresses typically contain more substance and less style than inaugural addresses (Campbell and Jamieson 2008)).[1] And finally, our approach allows us not only to assign statistical significance to the thematic content of these speeches, but also to employ correspondence analysis to depict spatially the shifting dimensionality in themes used by presidents. We therefore offer a systematic and statistically robust way to gauge the extent to which Reagan's rhetoric was distinct from other recent presidents.

Our findings offer strong evidence for Reagan's unique usage of the civil religion rhetoric: well over half (59%) of the discourse in his seminal speeches and 48% of the same in his State of the Union speeches focus on themes of civil religion. While other modern presidents invoke some of these same themes in their State of the Union speeches, none were as explicit or as extensive as Reagan, particularly in the usage of "God." Perhaps even more interesting is a secondary—and unexpected—finding from our correspondence analysis of the State of the Union speeches of all the modern presidents. We find what appears to be a distinct shift in presidential rhetoric from themes concerned with (1) *institutions*, to ones focused more on (2) *individuals, families, and children*. All presidents from Reagan onward have spoken more about individuals, families, and children; whereas their predecessors (with the exception of Lyndon Johnson) spoke more about institutions (e.g., relations between the executive and Congress, federalism). Whether Reagan was a rhetorical "pivot" in this respect may be debated; but it does raise implications for further research.

We begin in the next two sections with a short overview of civil religion rhetoric in presidential speeches and particularly that by Reagan. In the third section we explain our data and provide a brief description of our methodology. The fourth section presents the results from our textual analysis, placing Reagan's rhetoric in historical context to ascertain the extent to which his use of religious terms differs from other presidents, and the fifth section concludes. Our Appendix provides a fuller description of our methodology and software.

1. Moreover, inaugural addresses are sometimes denoted as ceremonial speaking, whereas State of the Union addresses tend to be more deliberative in content, and so the use of civil religion as a ceremonial feature in the latter may be unusual. (We thank an anonymous reviewer for pointing this out.)

Civil Religion Rhetoric in Presidential Speeches

Defining Civil Religion

Many authors have sought to capture the essence of the concept of civil religion (Albanese 1976; Bellah and Hammond 1980; Dunn 1984; Hughey 1983), but the precise definition can vary from author to author. Some point to the *unifying effect* that it is said to exert on a diverse American public, pointing to "the collection of beliefs, values, rites, ceremonies, and symbols which . . . provide [Americans] with an overarching sense of unity above and beyond all internal conflicts and differences" (Pierard and Linder 1988, 23). Others highlight the nation's *unique history* as giving rise to its identity as a *chosen nation*: "At the core of the rich and subtle concept of civil religion is the idea that a nation tries to understand its historical experience and national purpose in religious terms . . . so a civil religion reflects an attempt by citizens to imbue their nation with a transcendent value. The nation is recognized as a secular institution, yet one that is somehow touched by the hand of God" (Wald and Calhoun-Brown 2007, 54). Chidester organizes the various strands into three types of American civil religion (1988, 83-109): (1) *culture* religion, which includes the symbols, values, and ceremonies that serve to unify Americans—such as myths (heroic battles, founding fathers), doctrines (equality, liberty and justice), ritual holidays (Thanksgiving, Fourth of July, Memorial Day), patriotism and love of country as personal experiences, and key institutions that serve as carriers of civil religion (especially the public school system); (2) religious *nationalism*, whereby Americans are seen as the chosen people, and manifest destiny encapsulates the American moral mission to extend its political power geographically (including against the forces of communism during the Cold War); and (3) *transcendent* religion, where America's actual history and experience is judged and "measured against a higher law" or a set of sacred principles (Chidester 1988, 85). In a somewhat parallel attempt to define civil religion, Lejon (1988, 198) traces its roots to three sources, from which he then lists defining keywords: (1) Puritanism (chosen land, chosen people, mission, moralism, etc.); (2) Enlightenment (natural law, reason, moral teaching, democracy, etc.); and (3) American history (heroes, sacrifice, moral crusades, etc.).

As is evident from above, the concept of civil religion is multifaceted and complex. Its inherent lack of a clear and precise definition makes any test of it a challenge (Greeley 1989, 119). As Wald and Calhoun-Brown contend, the term "does not refer to any formal code of beliefs that is fully developed and authoritatively encapsulated by a single written document. Because there is no formal statement of it, the content of a civil religion has to be inferred from the speeches and writings of political leaders" (Wald and Calhoun-Brown 2007, 54)—and, in particular, those of the president. Hence, we do not accept at face value *any* of the definitions above as the single, definitive representation of civil religion. Rather, we provide an operationalization of the terms that we contend represent the common denominator in our reading and understanding of the civil religion literature. We list these terms and provide further explanation when we turn to our expected results (below).

Finally, within the larger context of presidential rhetoric, our focus here is on studies that have employed traditional content analysis or computerized methods to explore presidents' speeches from George Washington onward (some examples include (Bimes and Mulroy 2004; Ellis 1998; Foerstel-Branson 2009; Lim 2008; Tulis 1987)—and more narrowly on those that have revealed insights into the religious rhetoric of presidents. In her study of inaugural addresses from Washington to Reagan, Toolin (1983) (using traditional content analysis) found clear evidence of civil religion, but, as noted above, the use of inaugural addresses biased her sample in favor of finding such rhetoric. More recently Lim employs computer-assisted content analysis on both inaugural and State of the Union speeches, finding a significant increase in religious rhetoric. From his analysis of over 200 years of presidential speeches, Lim (2002) finds that presidential rhetoric has become more abstract in recent decades—and by abstract he means that presidents have increasingly invoked religious words in their rhetoric. The use of "God" in particular has become more prevalent in twentieth-century presidential rhetoric, especially during Ronald Reagan's presidency. Lim's use of computer-assisted content analysis to uncover patterns and anomalies in political texts reflects a rapidly growing attention to the role of ideas, arguments, and deliberation in politics. We build on Lim's findings by a similar but more extensive methodological approach to examine in greater depth the extent to which Reagan constitutes an outlier in the use of religion in presidential rhetoric.

Reagan and Civil Religion

Some critics of Reagan's use of religious rhetoric have pointed to his failure to attend religious services himself, his divorce, and his wife Nancy Reagan's fascination with astrology as evidence of the superficiality of this rhetoric. However, as noted above, the use of civil religious rhetoric by presidents is not presumed to reflect their own private religion, but rather some common thread that is said to be shared by the majority of Americans. Our concern here is *not* to discern whether Reagan's espousal of civil religious beliefs was sincere (as tributes to Reagan by the late Senator Edward Kennedy and Senator Norm Coleman [U.S. Congress, Joint Committee on Printing 2005, 233, 242] suggest) or mere lip-service (as others imply [Fowler, Hertzke, and Olson 1999, 119]), or whether Reagan even fully understood the context and historical foundations for his civil religious rhetoric (Edel 1987, 151-52). Our task in this article is more targeted: we seek to assign a numeric weight to the prevalence of civil religious rhetoric in Reagan's speeches and to compare this to the equivalent rhetoric by other modern presidents. The literature offers ample anecdotal evidence of Reagan's visionary rhetoric[2] but no systematic measure for its statistical significance. We thus seek to use computer-assisted content analysis to discern the extent to which Reagan constitutes an empirical outlier with respect to religious rhetoric.

2. Perhaps the most famous is Reagan's use of John Winthrop's vision of a "shining city for all the world upon the hill."

In a more critical vein, one might ask, even if Reagan may be established as a rhetorical outlier with respect to religion, what does this matter for real policy or outcomes? One response is that presidential rhetoric is indicative of "how the president perceives his political position"—that is, who he aligns with and who he opposes (Bimes and Mulroy 2004, 141) and therefore helps us to understand his partners and priorities. Second, political theorists have rightly noted that even if political speeches are not in themselves causally effective, they may prove useful for legitimizing programs and projects (Weale 2010, 270). Skinner shrewdly outlined ways in which the "innovating ideologist" might "manipulate an existing normative vocabulary in such a way as to legitimate . . . new and untoward courses of action," including the introduction of new or previously neutral terms into the political discourse and applying these to some course of action which the author favors (1974, 296). Reagan's rhetoric may, then, help us to understand better the ways in which he sought to legitimate his conservative ideology and policy priorities.

Data and Methodology

Data

Because scholars in presidential rhetoric and communication agree that presidents are most likely to articulate their primary and core beliefs in their seminal speeches,[3] we begin our analysis with these speeches for Reagan.[4] We have no *a priori* expectations concerning the empirical weights of the thematic content of Reagan's speeches—nor that of any other president in our sample. We do, however, contend that a distinct pattern of association between each president and his core beliefs should emerge. We maintain that different themes of rhetoric that use different vocabulary will result in an observed word distribution that deviates systematically from one where the words are independent of each other.

Our data for Reagan are drawn from the Ronald Reagan Presidential Library web site (http://www.reagan.utexas.edu/archives/speeches), and his seminal speeches are selected from his time as president (January 1981 to January 1989). The Reagan Library web site posts a total of 94 "major" speeches during the Reagan presidency but does not distinguish the seminal ones. However, of the 94, 10 are considered Reagan's most important speeches to the nation.[5] To test for the possibility that Reagan tailored his

3. We are grateful to Terri Bimes and Mary Stuckey for their recommendations in tailoring our sample of speeches.

4. Moreover, a sample that includes nonseminal speeches invites criticism, as it would invariably require more subjective judgment in the selection process.

5. Two of these are inaugural speeches (January 1981, January 1985); seven are State of the Union (only three were given in his first term, in January 1982, January 1983, and January 1984; and four were given in his second term, in February 1985, February 1986, January 1987, and January 1988); and one is his farewell speech (January 1989). These speeches are considered by scholars in presidential rhetoric to be the most important to the Office of the President, since they "attempt to influence coming legislative programs, inaugurals and farewells by creating a view of the country and its future, State of the Union addresses by setting a legislative agenda and making specific recommendations. . . . [They] include public mediations on

religious rhetoric to more receptive audiences, we add to the sample Reagan's eight annual keynote speeches (1981-88) to the Conservative Political Action Conference. As will be seen, our analysis allows us to see whether there is a significant and systematic difference between Reagan's speeches to conservative audiences versus those to the broader American public.

State of the Union addresses from Woodrow Wilson to Barack Obama were obtained from the American Presidency Project (http://www.presidency.ucsb.edu/)[6] and divided into two text files—the pre– and post–World War II presidents (i.e., Woodrow Wilson to Franklin Delano Roosevelt [FDR], and Harry Truman to Barack Obama).[7] For simplicity, our only identifying tag is the name of the president.

We structure the data into four text files: (1) Reagan's seminal speeches, (2) Reagan's State of the Union speeches, (3) the State of the Union speeches for Wilson to FDR, and (4) these same set of speeches for Truman to Obama. The text files are structured so each speech constitutes a "case," and each is identified (or "tagged") with identifying characteristics—the name of the president and for Reagan's seminal speeches, the nature of the speech (Inaugural, State of the Union, Republican Party [CPAC], and Farewell).

A Typology of Expected Results

From our brief overview, we might anticipate finding some evidence for Reagan's usage of inspirational language. However, as we noted earlier, the lack of a clear definition or set of terms by which to test for civil religion raises a difficulty as to how to differentiate between civil religion and noncivil religion rhetoric. Our approach is automated and so there is no precoding of the text (a topic we discuss in our Appendix). It is, however, useful to specify in advance certain terms that appear to comprise the core understanding of civil religion. Earlier we noted the absence of a clear, agreed-upon definition of civil religion, but there are some terms around which most authors converge. These terms are not exhaustive (or listed in any particular order) but instead provide the fundamental basics for civil religion rhetoric: *mission, sacrifice, destiny, chosen, freedom, divine/providence/spirit/God, America as an international example*. To the extent that our analysis finds systematic evidence of this rhetoric, our article lends a precise empirical measure to what have often been presented as vague generalizations. We do not maintain that our operationalization of civil religion is the only way to measure the term empirically or that it does so perfectly. For instance, one might conceptualize American Exceptionalism as a closely related concept—and one on which presidents might speak without reference to religious rhetoric. We contend that there is a fundamental overlap between basic societal secular myths and civil religion, so that some of our key terms (mission, sacrifice, destiny) may apply to both categories. However, the *combined* usage of

---

national values and identity" (Campbell and Jamieson 1990, 212). Moreover, (1) the Constitution requires the president to give the State of the Union to Congress (and since Wilson this has been in the form of a speech); and (2) unlike most other speeches, it is delivered to a national audience and receives widespread media coverage. Inaugural and farewell addresses are also prominent in capturing national media coverage.

6. We are very grateful to John Woolley for sending these files to us in a more manageable format.

7. Our data cover up to 2010.

religious rhetoric and the remaining terms is unique to civil religion—and given our methodology's effective use of co-occurrence analysis (which we discuss below and in our Appendix), is something that we seek to examine empirically. In short, there are a variety of ways that presidents might talk about the American mission, and civil religion is just one of these.

Our findings may be organized according to a schema that is typical of the results of statistical text analysis, namely, that the results usually fall into one of five categories: the trivial, the classic, the unexpected, the artifact, and the residue.

The trivial are those that are so obvious as to be uninteresting (for example, that the discourse of political leaders contains political terms with significantly high frequency). Statistical software, having no knowledge of the world outside of the text will invariably convey information that is so commonsensical that it is not worth mentioning.

The classic result is what experts already know—or think they know. In our case, that Reagan expressed civil religion rhetoric is known anecdotally but not empirically. Hence, is it satisfying to verify that these classic results are grounded in empirical evidence, but the software also enables quantification, which is especially informative for comparing the rhetoric of different presidents.

Generally speaking, statistical text analysis provides greatest value in revealing unexpected results. Because they are naïve, software have little preconceptions and may uncover counterintuitive or odd patterns, while a researcher will invariably be limited by her preconceptions or confirmation biases (Nickerson 1998).

Artifacts are those results that come from the specific biases of the algorithms. For example, all software based on the calculation of co-occurrence will be especially sensitive to repeated chains, ideographs, or idiomatic expressions that create clusters: for example, "United States of America" or "God bless America." These must be spotted and corrected, for example by transforming the term into a single item (in our software, a single term would be joined by underscores, e.g., "God_bless_America"). There are also more subtle artifacts; in general, the analyst must use her knowledge of the algorithms to check whether unexpected results are due to some artifact.

Finally, the residue is what the analyst is unable to interpret. This may come from the analyst's limitations or from the fact that statistical analysis only sheds light on strong regularities, and therefore there is always a residue that will just be "noise." A careful analyst will be aware of the limitations of the statistical approach and therefore (1) not try to interpret everything and (2) be tolerant to the fact that some results may not fit ideally in the analysis.

Methodology: Computer-Assisted Content Analysis

The use of text-mining or text analysis software has proliferated in recent years, not least of which in the academic literature. A survey of these software lies outside the scope of this article; here, we confine our discussion to computer-assisted content analysis.

One form of automated content analysis is topic modeling, where the task is automatically to classify the contents of documents into "topics." Each topic is understood to comprise a distribution over a fixed vocabulary of words or terms, and each

document exhibits any number of topics in different proportions (Blei and Lafferty 2009). The basic idea is that words are indicative of topical content, and the task is to map the words into topics using a specified parametric form. These models are useful for exploring and cataloguing vast digital libraries (Blei and Lafferty 2006, 2009) or for categorizing a large number of speeches on a variety of subjects, where very little substantive knowledge of the subjects themselves is required (Quinn et al. 2010).

We do not adopt a topic model, since our approach to content analysis assumes that speakers or authors of textual data convey meaning in a more thematic fashion, and so it is not just the words that help to classify content but also the context in which the words appear. Rather than conceptualizing words in a univariate distributional pattern (e.g., as in topic modeling), a thematic approach examines the bivariate associations between words and sets of words in order to map out so-called lexical worlds (or concept clouds (i.e., concepts in which a given set of words *co-occur*) and the relationships between lexical worlds within a single corpus. Hence, our methodology allows us to capture the context and meaning of themes because it does not simply classify the contents of documents into "topics" based upon a univariate distributional pattern. To illustrate, in a topic model, each speech by a president might constitute a document, and then each document assigned to a topic (and just one topic) based on the largest proportion of words in the speech. Let us say that a given speech contains four components (topics), each with shares of 40%, 30%, 20% and 10% (and assume that 100% of the speech is successfully classified into topics). The whole of the speech would be allocated to the first topic, given that its share is the largest. Hence, 60% of the speech would be misallocated, since each speech is allocated to just one topic.

Our approach *does not do this*. Instead, we assume that speakers convey meaning in a more thematic fashion, and so it is not just the words that help to classify content, but also the *context* in which the words appear (e.g., it captures whether the speaker uses religious terms *together* with other words that signify American Exceptionalism).

Our software (Alceste) considers the text as a large matrix of co-occurrences between lexical forms and processes it with multivariate techniques. A key feature of Alceste is that it can be used to identify the speakers' tendency to articulate particular ideas and arguments—ideas and arguments that can then be correlated with characteristics of the speaker (e.g., in political texts—the name of speaker, party affiliation, constituency characteristics, and so on).

In our Appendix, we provide a more detailed description of the algorithms and their rationale, but in brief, Alceste operates in four steps: it parses the vocabulary (step A); it transforms the corpus into a sequence of Elementary Context Units (ECUs) containing words (or more exactly stemmed words or "lexemes") and operates a descending classification which produce stable classes of these ECUs, leaving what does not fit in these classes "unclassified" (step B); it operates a series of statistical characterizations of the classes (typical words, typical sentences, crossing variables, providing χ^2 values, etc.) (step C), which enable the analyst to operate interpretation (step D). The interpretation consists in attributing meaning to the "lexical world" that is latent in each class based on these statistical results. The software provides a number of tools for the researcher to interpret each class, and two tools are particularly useful—the characteristic words and

the characteristic phrases.[8] Both are ranked in order of χ^2 significance, to allow a clearer understanding of the terms and phrases which predominate in each class.

Results from Textual Analysis

Reagan

Table 1 provides summaries of the basic statistics from Alceste for Reagan's seminal and State of the Union speeches. Turning first to the seminal speeches, we observe a total word count of 71,335.[9] The passive variables[10] (or tags) define characteristics of each speech or "case," and these include Reagan's name, the nature and date of the speech. Each speech within each corpus constitutes a sampling unit and is called an Initial Context Unit (ICU) by the software. These ICUs are cut into ECUs, which are the basic elements of the classification process. As a measure of goodness of fit, we observe that 1240 ECUs are classified, giving an overall classification rate of 72%. Recalling our typology of results, 28% of the seminal speeches corpus may thus be considered as residue.

The bottom two rows in Table 1 indicate the number of classes identified and the size of each class (as measured by the percentage of the total ECUs classified within each). In total, four classes are identified for the seminal speeches: *civil religion, big government/ fiscal policy, domestic economic growth*, and *Soviet threat*.

TABLE 1
Basic Statistics for Reagan's Seminal and State of Union Speeches, in Separate Analyses

	(a.) Seminal speeches	(b.) State of Union speeches
Total Word Count	71,335	32,170
Unique Words Analyzed	30,153	13,477
Number of speeches	18	7
Classified E.C.U.s	1240 (=72% of the retained E.C.U.)	660 (=83% of the retained E.C.U.)
Lexical Classes	4	5
Distribution of Classes (%)	1 58.5% Civil Religion	1. 17.3% Soviet Threat / National Security
	2 19.7% Big Government/Fiscal Policy	2. 11.8% Federal Budgetary Reform
	3 8.7% Domestic Economic Growth	3. 12.0% Other Policy Reform Initiatives
	4 13.2% Soviet Threat	4. 10.8% Macroeconomic Policy
		5. 48.2% Civil Religion

8. See (Lahlou 1995b) for a detailed description of the interpretation procedure and its theoretical basis.

9. Plurals and conjugation endings are reduced to a single form and nonce words are eliminated from the analysis. This leaves a smaller word count, which is analyzed by the program.

10. These are deemed "passive," as they do not contribute to either the calculation of the word classes or the factors in the correspondence analysis.

While the assigned class labels may seem straightforward, it is important to clarify that these are not automatically given by the program. The output provides the researcher with a number of different tools for conceptualizing the content of classes. As noted earlier, of the many tools, two are particularly useful—characteristic words and characteristic ECUs.[11] The most characteristic function words for each class (ranked in terms of their χ^2 statistical significance, with the minimum χ^2 value for selection set at 4.2 for this analysis, with one df[12]), provide an indication of the theme or frame of argument that unifies a class. The most characteristic words for each class are those with the highest rankings. To avoid repetition, we postpone our illustration of the labeling of classes to the State of the Union speeches and for now, we take the labels as given.

The most striking observation from the classification of themes in Reagan's seminal speeches is the very large percentage accounted for by *civil religion* rhetoric. We anticipated some evidence of this rhetoric, but we could not have foreseen its impressive weight (58.5%) relative to Reagan's other thematic priorities.

Even so, Reagan's focus on civil religion—which is about three times the size of the second largest theme of *big government*—appears extraordinary and deserves further investigation. To delve deeper, we use the *Tri-Croisé* or Cross-Data analysis in Alceste, which crosses a tag (name of speaker, etc.) or a single word with the entire text and identifies the strongest statistical associations between the specified tag or word and other words and phrases in the text.[13] (Simply put, this holds constant the specified term, allowing all else to vary.) From the *civil religion* class in Reagan's seminal speeches corpus, the top two characteristic words are *God* and *conservat*+.[14] Using *God* in the *Tri-Croisé*, we find the strongest association with the word *bless*, and examining the most significant ECUs, it is clear that the most frequent usage of *God* by Reagan (i.e., 26 out of a total of 76) was in the phrases "God bless America" or "God bless you." As such phrases might reasonably be considered vacuous, we transform the 26 usages of the individual words "God bless" into the term "God_bless," thereby controlling for it relative to more substantive usage of "God." We reanalyze Reagan's seminal speeches and find that the previous *civil religion* class divides into two distinct new classes.[15] The first new class exhibits the top characteristic words of *conservat*+, *think, movement*+, *politic*+, *want*+, *agenda, conference, liberal*+, and *elect*+—which, together with the ECUs, suggest a theme centering around conservatism as a political ideology. The second new class lists the top characteristic words of

---

11. The standard report lists the top 20 ECUs for each class, ranked by chi-square association. However, a separate file is produced that lists all the ECUs for each class, where the default cutoff for selection is zero.

12. This minimum value for word selection within Alceste varies from 2.13 to 20, with smaller text files tending toward the lower threshold and larger ones toward the high threshold. The basic rule of thumb with Alceste is (as with any statistical analysis)—the more data, the easier it is to attain statistical significance (hence larger text files have to attain a higher threshold to be statistically significant).

13. For a good example of this technique applied to parliamentary debates see (Bicquelet 2009).

14. The "+" indicates the lemma version (or base form) of word, so that conservat+ may include conservative, conservatives, conservatism, and so on.

15. Notably, the percentage of ECUs classified also improves to 80%.

God+, *dream, faith+, love+, children+, court+, neighborhood+, school+, courage, heart+, unborn, citizen+, value+, abortion+, life, spirit+, prayer+, land+, teach, compassion+*, famil+, and *societ+*. Top ECUs, or phrases, (with characteristic words in bold) further illustrate the content of this class:

> if you take away the **dream**, you take away the **power** of the **spirit**. If you take away the **belief** in a **greater** future, you cannot **explain America**—that we're a **people** who **believed** there was a **promised land**; we we're a **people** who **believed** we were **chosen** by **God** to create a **greater world**.

> We **advance** the **promise** of **opportunity** every time we speak out on behalf of lower tax rates, **freer markets, sound currencies around** the **world**. We strengthen the **family** of **freedom** every time we work with allies and come to the **aid** of **friends under** siege. And we can enlarge the **family** of **free nations** if we will defend the unalienable **rights** of all **God's children** to **follow** their **dreams**.

> **America** was **founded** by **people** who **believed** that **God** was their rock of **safety**. He is ours. I **recognize** we must be cautious in **claiming** that **God** is on our side, but I **think** it's all **right** to keep asking if we're on His side.

> and while I'm on this **subject, each day** your members observe a 200 year old **tradition** meant to signify **America** is one **nation under God**. I must ask: if you can begin your **day** with a member of the clergy **standing right** here leading you in **prayer**, then why **can't freedom** to **acknowledge God** be **enjoyed** again by **children** in every schoolroom across this **land**?

By stripping away the "God bless" catch phrases, and separating out the more overt conservative political ideological content of the civil religion theme, we contend that this new thematic class (which constitutes 21% of the classified ECUs) is not simply a ploy to conservative audiences, as the discourse is not targeted at specifically Republican audiences but rather is spread across all his seminal speeches.[16] From the characteristic words and phrases, we see Reagan's emphasis on *God* and *freedom*, as well as expressions of America as a *chosen people* and as an *international example*—all of which fit precisely within the core rhetoric that we have identified for civil religion. Reagan's unique contribution to this rhetoric is the added "visionary" component—namely, the reliance on "dreams" and "greatness"—as well as his focus on life (of the unborn) and freedom for children to pray in school.

Turning to Reagan's State of the Union speeches (where we control for the God_ bless phrase), we see once again his focus on civil religion rhetoric, with this class comprising 48.2% of the overall ECUs. Inasmuch as the State of the Union speeches comprise our basis of comparison with other presidents, we expand our analysis here to illustrate our approach to labeling each of the classes. From Table 1, Reagan's State of the Union speeches exhibit five classes: *Soviet threat/national security; federal budgetary reform; other policy reform initiatives; macroeconomic policy;* and *civil religion*.

Table 2 presents the most representative words and phrases for each of the five classes in these speeches. For each class, we present the top dozen characteristic words,

---

16. Statistical significant tags for this class span across his inaugural, State of the Union, and CPAC speeches.

TABLE 2

Characteristic Words and Phrases for Classes in Reagan's State of Union Speeches

Class (size)	Top Characteristic words	Top 3 Characteristic Phrases (ECUs) (Characteristic words in bold)
1—Soviet Threat / National Security (17.3%)	Soviet+ arms nuclear negotiat+ allie+ central+ Nicaragua+ democrat+ support+ fighters agreement Afghanistan	. . . the protection of our national **security** has **required** that we **undertake** a substantial program to enhance our **military forces**. We have not **neglected** to **strengthen** our traditional **alliances** in **Europe** and **Asia**, or to develop **key relationships** with our **partners** in the **Middle East** and other countries. Your **vote** this spring on the Peacekeeper **Missile** will be a **critical** test of our **resolve** to **maintain** the **strength** we need and **move toward mutual** and **verifiable arms** reductions. Meanwhile, we're working for reduction of **arms** and **military** activities, as I announced in my address to the nation last November 18th. We have proposed to the **Soviet Union** a far-reaching **agenda** for **mutual** reduction of **military forces** and have already initiated **negotiations** with them in **Geneva** on intermediate-range **nuclear forces**.
2—Federal Budgetary Reform (11.8%)	federal+ billion+ program+ budget+ spend social needy stamp+ total+ saving+ dollar+ over	. . . the **savings** we **propose** in **entitlement programs** will **total** some $63 **billion over** 4 **years** and will, without affecting **Social** Security, go a **long** way toward bringing **federal spending under control**. **National** security is government's first **responsibility**; so in past **years** defense spending took about half the **federal budget**. **Today** it **takes** less than a **third**. We've already reduced our planned **defense** expenditures by nearly a **hundred billion dollars over** the past 4 **years** and reduced projected **spending** again this **year**. Virtually every American who shops in a **local** supermarket is aware of the daily **abuses** that **take** place in the **food stamp program**, which has **grown** by 16, 000 percent in the **last** 15 **years**.
3—Other Policy Reform Initiatives (12.0%)	private sector act+ train+ regulatory help+ area+ reform+ clean+ quality+ competitiveness teacher+	. . . and already, **communities** are **implementing** the commission's recommendations. **Schools** are **reporting** progress in **math** and reading **skills**. But we must do more **to restore discipline to schools**; and we must **encourage** the **teaching** of **new basics, reward teachers** of **merit, enforce tougher standards**, and put our **parents** back in charge. . . . **continue to** support the **job training partnership act**, which has a nearly two-thirds **job** placement rate. **Credits** in **education** and **health care** vouchers will **help working families** shop for **services** that they need. The **Department** of **Transportation** will **help** an expendable **launch services** industry **to** get off the ground. We'll **soon implement** a number of **executive initiatives, develop proposals to** ease **regulatory** constraints, and, with NASA's **help, promote private sector** investment in space.

TABLE 2 *Continued*

Class (size)	Top Characteristic words	Top 3 Characteristic Phrases (ECUs) (Characteristic words in bold)
4—Macroeconomic Policy (10.8%)	rate+ inflation+ taxe+ tax deficit+ interest+ percent+ cut unemploy+ revenues high+ government+	**We've brought inflation down faster** than we **thought** we could, and in doing this, **we've** deprived **government** of those hidden **revenues** that occur when **inflation** pushes people **into higher income tax** brackets. We who are in **government** must take the **lead** in restoring the **economy**. And here all that time, **I thought** you were **reading** the paper. The single **thing** the single **thing** (sic) that can **start** the wheels of **industry turning** again is further **reduction** of **interest rates.** **Together**, we passed the first across-the-board tax reduction for **everyone** since the Kennedy **tax cuts**. Next year, **tax rates** will be **indexed** so **inflation** can't **push** people **into higher** brackets when they get cost-of-living **pay raises.**
5—Civil Religion (48.2%)	nation+ America+ dream hope+ great+ heart+ tonight spirit+ children+ let God land	. . . memory of our **great men** it means those **young Americans** will **find** a city of **hope** in a **land** that is **free.** We can be **proud** that for them and for us, as those **lights** along the Potomac are still seen this **night** signalling as they have for nearly two **centuries** and as we **pray God** they always will, I **see America** in the crimson **light** of a **rising sun** fresh from the burning, **creative hand** of God. I **see great days** ahead for **men** and women of will and **vision.** Some questioned whether we had the will to **defend** peace and **freedom**. But **America** is too **great** for **small dreams.** There was a hunger in the **land** for a **spiritual** revival; if you will, a crusade for **renewal**. The **American people said: let** us **look** to the **future** with **confidence**, both at **home** and **abroad.**

and the three top characteristic phrases. For Class 1, the key words *Soviet+*, *arms, nuclear, negotiate+*, *allie+*, *central+*, *Nicaragua+*, *democrat+*, *support+*, *fighters, agreement*, and *Afghanistan*, combined with the contextual references to U.S. national security, containing the Soviet Union (through alliances and arms reductions) all point to a theme of *Soviet Threat/National Security*.

Following this same mode of interpretation, Class 2 focuses on reforms to the *federal budget* (Social Security, defense spending, food stamps); Class 3 focuses on other *policy reform initiatives* (education, employment, private sector initiatives); and Class 4 discusses *macroeconomic policy* (inflation rate, fiscal policy, unemployment). The final class—*civil religion*—is, of course, our primary concern. From both the characteristic words (*nation+*, *America+*, *dream, hope+*, *great+*, *heart+*, *spirit+*, *children+*, *God*) and phrases, we again see evidence of the core civil religion rhetoric (*God* and *freedom*, America as a *chosen people* and as an *international example)* as well as his unique emphasis on dreams, greatness, and hope for a brighter future (though, in contrast to his seminal speeches, an absence of "abortion" in the characteristic words and phrases).

Reagan Relative to Other Presidents

Reagan's rhetoric both in his seminal and State of the Union speeches exhibits clear evidence of the civil religion discourse. But how unique is this from other modern presidents? If we take on board Lim's (2002) empirical finding that the use of "God" became particularly more frequent with Reagan's presidency, there is reason to suspect a shift in presidential rhetoric, beginning with Reagan. That would imply, however, a similarity between Reagan and subsequent presidents. So even if Reagan's religious rhetoric was different from previous presidents, it may be less so from subsequent presidents.

To test Reagan's religious outlier status relative to other presidents, we compare Reagan's State of the Union addresses with all the presidents from Woodrow Wilson to Barack Obama. We seek to extend Lim's finding, but in so doing, our examination complements other recent applications of computerized content analysis to study presidential rhetoric—e.g., the use of the DICTION software, which uses predefined dictionaries to generate composite variables (optimism, blame, denial, hardship, etc.) (see (Bligh, Kohles, and Meindl 2004; Foerstel-Branson 2009). Rather than create composite variables, our software generates thematic classes from which further exploration may be pursued (as we show above). Other recent studies of presidential rhetoric have retained some element of traditional content analysis, blending this with computer searches for key words or phrases (Bimes and Mulroy 2004). Lim's extension of his earlier article employs a novel tool—the Flesch Readability Score (which measures sentence length and number of syllables)—to explore the density and substance of presidential speeches (Lim 2008). He finds that presidential rhetoric has declined in quality, particularly in terms of logical argument. For both his book and article, he uses the software General Inquirer (GI), which relies on dictionaries of tags for categories of words (e.g., the tag "feel" may include such words as "aloof" and "apathy") to identify patterns of occurrence.[17] Our approach, as described earlier and in our Appendix, has no preset categories but instead relies upon co-occurrence analysis to identify lexical worlds or thematic classes.

As noted earlier, we analyze the pre– and post–World War II presidents separately (i.e., Wilson to FDR, and Truman to Obama {as of 2010}). For simplicity, our only identifying tag is the name of the president. In anticipation of the "God bless" term, we again create the term "God_bless" so as to isolate this from potentially less vacuous uses of "God." Finally, in order to simplify the base-line of comparison for civil religion rhetoric, we focus predominantly on the use of "God."

For both time periods, we report both the thematic classes and those presidents whose name tag obtains statistical significance, along with the weight of each class (as defined by the percentage of ECUs classified in each class). It is not our intent, nor do we have the space here, to examine each thematic class. Rather, our goal is to determine the extent to which other presidents used civil religion rhetoric—and particularly the term "God"—in their State of the Union addresses. In neither time period did we find a

---

17. For a description of the categories and dictionaries, see http://www.wjh.harvard.edu/~inquirer/homecat.htm (accessed April 13, 2012).

TABLE 3
Basic Statistics for State of Union Addresses, Wilson to Roosevelt (FDR)

Total Word Count	170,708
Unique Words Analyzed	72,148
Passive Variables (Tagged Indicators)	5
I.C.U.s (=number of cases)	5
Classified E.C.U.s	2920 (=71% of the retained E.C.U.)
Lexical Classes	6

	Theme of Class	Significant Tag	Weight of Class (%)
	1. *Effect of Depression on Agriculture, Taxes*	Hoover	30.5
	2. *Executive-Congressional Relations, especially on Regulation (Utilities, Prohibition, Shipping)*	Hoover, Coolidge	26.2
	3. *Value of Democracy, Freedom, Life, Economic Security*	Roosevelt (FDR)	12.3
	4. *National Defense, Honoring Fallen Soldiers*	Wilson, Roosevelt (FDR)	12.8
	5. *Foreign Relations, Int'l Peace Conference*	Coolidge	10.6
	6. *National Infrastructure (Waterways, Navigation)*	Coolidge	7.6

thematic class centered around God (or the divine/providence/spirit). Nor did we find the term "God bless" in *any* of the pre–World War II addresses. The first usage of "God bless" was by Truman in his final State of the Union address in 1953, and he used the term just once. The next time a modern president used the term in his State of the Union address was Gerald Ford in 1977, and like Truman he used it just once. We find a striking shift with Reagan, who adopted the phrase "God bless you and God bless America" as his standard closing, and thereby used the term 12 times in his addresses. After Reagan's example, every president followed suit, with George Bush using the term eight times, Bill Clinton 12 times, and G. W. Bush six times (although interestingly, he preferred a simple "God bless" with no subsequent "you" or "America"). As of 2010, Obama used the term four times.

This finding, although intriguing, may simply reflect an acquired window-dressing for State of the Union closings. Nonetheless, we examine the thematic classes in Tables 3 and 4 to gain a better understanding of the potential use of religious terms. In the early period (Table 3) we find one class in which the term "God" appears among the characteristic words, although ranked as thirty-first in significance, well below the top words: "right+," "world+," "democracy+," "secur+," "nation+," and so on. The theme of this class is that of *human values*—particularly the values of *life, liberty, and freedom*—along with discourse on the *human spirit* and importance of *faith*. Name tags for the presidents

TABLE 4
Basic Statistics for State of Union Addresses, Truman to Obama 2010

Total Word Count	421,783		
Unique Words Analyzed	197,611		
Passive Variables (Tagged Indicators)	12		
I.C.U.s (= number of cases)	12		
Classified E.C.U.s	5724 (=77% of the retained E.C.U.)		
Lexical Classes	7		
	Theme of Class	*Significant Tag*	*Weight of Class (%)*
	1. *Proud Americans, Moments in History (Motivational)*	Clinton, Bush, Reagan	13.2
	2. *Executive-Congressional Relations & Coordination; Federalism*	Nixon, Eisenhower, Carter	21.1
	3. *Human Values (life, liberty, freedom); Human Spirit; Faith*	Truman	10.0
	4. *Education & Welfare*	Clinton	12.4
	5. *Deficit Reduction, Taxes, Inflation & Unemployment*	Reagan, Obama, Clinton	10.6
	6. *Agriculture & Business*	Truman	10.1
	7. *Foreign Policy (Cold War, alliances)*	Carter, Kennedy, Eisenhower	22.6

may obtain statistical significance according to their χ^2 value (with 1 df): over 3.84 for 10%, over 6.63 for 5%, and over 10.8 for 1%. For this class, one president is statistically significant: FDR with a χ^2 value of 232.

Turning to the Truman-Obama period (Table 4), there are two classes of interest. In Class 3, we find language that closely resembles the *Human Values; Human Spirit: Faith* class dominated by FDR. In this later period, it is Truman who is statistically significant ($\chi^2 = 29$), although Richard Nixon and G. W. Bush obtain marginal significance at the 5% level ($\chi^2 = 8$ and 7). For this class, the term "God" obtains significance,[18] but is only ranked twenty-second—well below the top terms of "freedom+," "spirit+," "faith+," "life+," and "liber." A second class contains vocabulary that is motivational, stressing "proud Americans," critical moments in history, and so on. In this class, the term "God bless" is a characteristic term but is ranked at just eighty-second. A top ECU from this class (which is from Clinton's 1997 address) illustrates the motivational theme:

> I **want** to **say** a **word** about two other **Americans** who **show** us how. Congressman Frank Tejeda was buried yesterday, a **proud American** whose family **came** from Mexico. He was only 51 years old. He was awarded the Silver Star, the Bronze Star, and the Purple **Heart fighting** for his **country** in **Vietnam**.

18. The software assigns χ^2 significance for characteristic words and phrases in order to assign a rank order list for each. However, unlike the tagged indicators, the words and phrases are not exogenous to the corpus and so we cannot interpret the χ^2 values according to the standard criteria for statistical significance. Hence, we simply report their rank order.

For this class, Clinton is overwhelmingly dominant ($\chi^2 = 180$), although G. Bush, Reagan, and Obama also obtain significance ($\chi^2 = 78$, 49 and 34, respectively).

From these findings, it is evident that five presidents (aside from Reagan) exhibit rhetoric in their State of the Union addresses that feature religious terms with statistical significance—FDR, Truman, G. Bush, Clinton, and Obama—and consequently these deserve closer scrutiny. Recognizing that Obama's term is unfinished, however, we exclude him from closer examination. We do, however, add G. W. Bush as a potentially relevant candidate for religious discourse, given his status as an evangelical Christian. We examine six corpora, one for each of the six presidents' State of the Union addresses. While space does not allow a full presentation of these results, we provide a summary overview for these six presidents in Table 5. Aside from Reagan, none of the other presidents obtain a thematic class that could be construed in any way as focusing on God/the divine/providence/spirit. Other presidents do, however, express some of the other core expressions of civil religion—namely, *freedom*, and *America as an international example* (especially against communism for Truman or against totalitarianism for Bush). But the norm among these presidents is to speak of American values more generally and religion only in the context of religious freedom.

In Table 5, we list the class for each of our selected presidents that most closely relates to civil religion. In column 3 we indicate the relative size of this class, as given by the percentage of ECUs classified within that class. Column 4 ranks the term "God" according to its significance value, relative to the other characteristic words for that class, and column 5 ranks the same for the term "God bless," where relevant. The final column gives two of the top phrases for each class (adding one further example for Reagan, to illustrate the unique language). It is evident both from the size of Reagan's civil religion thematic class (nearly half, or 48%) and the high ranking for the term "God" that his rhetoric was indeed unique relative to that of other presidents who might be expected to score high in use of religious terms. While previous narratives, biographies, and analyses of Reagan's rhetoric may have observed its unique qualities, not all have agreed on precisely which phrases were significant nor have they offered robust statistical foundations for their observations. In sum, we offer a statistically robust analysis of the extent to which Reagan used religious terms with far greater frequency than either his predecessors or successors. We also find that while presidents after Reagan adopted his "God bless" closing for State of the Union addresses, this is arguably rather superficial rhetoric as it is not supplemented with any civil religion thematic foundations.

Correspondence Analysis: A Spatial Depiction of State of the Union Addresses

We have, thus far, gained some understanding of the main themes in the State of the Union addresses, but not the relationships between these thematic classes. Our approach facilitates this by cross-tabulating classes and words in their root form in order to create a matrix that can then be subjected to factor correspondence analysis (Greenacre

TABLE 5
Comparison of Civil Religion or American Values Themes in State of Union Address Across Key Presidents

President	Theme of Class	Share of Classified ECUs (%)	Rank of "God" as Characteristic Word	Rank of "God bless" as Characteristic Word (Phrase)	Typical Phrases (from top 3)
Reagan	**Civil Religion**	48.2	11th	102nd	[see Table 2]
FD Roosevelt	Economic Security, Prosperity & Freedoms of Speech, Religion, Press	7.9	15th	Not significant	When you talk with our young men and women, you will find that with the opportunity for employment they want assurance against the evils of all major economic hazards—assurance that will extend form the cradle to the grave. I said then, and I say now, that these economic truths represent a second Bill of Rights under which a new basis of security and prosperity can be established for all—regardless of station, race, or creed.
Truman	Ideals of Freedom, Democracy, Peace—and Spreading these throughout the World (against Communism)	39.6	95th	Not significant	Let us all stand together as Americans. Let us stand together with all men everywhere who believe in human liberty. Peace is precious to us. It is the way of life we strive for with all the strength and wisdom we possess. But more precious than peace are freedom and justice. If we maintain and strengthen our cherished ideals, and if we share our great bounty with war-stricken people over the world, then the faith of our citizens in freedom and democracy will be spread over the whole earth and free men everywhere will share our devotion to those ideals.
Bush	Ideals of Freedom, Democracy (against Totalitarianism)	21.9	Not significant	Not significant	For two centuries, we've done the hard work of freedom. And tonight, we lead the world in facing down a threat to decency and humanity. What is at stake is more than one small country; it is a big idea: a new world order, where diverse nations are drawn together in common cause to achieve the universal aspirations of mankind—peace and security, freedom, and the rule of law. Still, we must recognize an unfortunate fact: in many regions of the world tonight, the reality is conflict, not peace. Enduring animosities and opposing interests remain. And thus, the cause of peace must be served by an America strong enough and sure enough to defend our interests and our ideals.
Clinton	Americans Sharing Common Future	21.5	Not significant	33rd	We Americans have forged our identity, our very union, from every point of view and every point on the planet, every different opinion. But we must be bound together by a faith more powerful than any doctrine that divides us, by our belief in progress, our love of liberty, and our relentless search for common ground. . . . an America that has continued to rise through every age against every challenge, a people of great works and greater possibilities, who have always, always found the wisdom and strength to come together as one nation.
GW Bush	War on Terror	55.2	Not significant	115th	Hezbollah terrorists, with support from Syria and Iran, sowed conflict in the region and are seeking to undermine Lebanon's legitimately elected government. In Afghanistan, Taliban and Al Qaeda fighters tried to regain power by regrouping and engaging Afghan and NATO forces. In Iraq, Al Qaeda and other Sunni extremists blew up one of the most sacred places in Shia Islam, the Golden Mosque of Samarra. States like these, and their terrorist allies, constitute an Axis of Evil, arming to threaten the peace of the world. By seeking weapons of mass destruction, these regimes pose a grave and growing danger.

and Hastie 1987, 437-47; see also Greenacre 1993).[19] In this way, we obtain a spatial representation of the relations between the classes. The positions of the points is contingent on correlations rather than coordinates (Reinert 1998, 45), where distance reflects the degree of co-occurrence.[20] With respect to the axes, correspondence analysis aims to account for a maximum amount of association[21] along the first (horizontal) axis. The second (vertical) axis seeks to account for a maximum of the remaining association, and so on. Hence, the total association is divided into components along principal axes. The resulting map provides a means for transforming numerical information into pictorial form. It provides a framework for the user to formulate her own interpretations rather than providing clear-cut conclusions.[22]

Figures 1 through 3 present maps of the correspondence analysis of the classes and name tags for the State of the Union addresses for Reagan alone, presidents Wilson to FDR, and Truman to Obama, respectively. The classes (and name tags, for Figures 2 and 3) are represented on a single graph—where distance between a class and a tag (or between two classes) reflects the degree of association.

Beneath the correspondence maps are the percentage associations for each factor, along with the cumulative for the two. Hence, a two-dimensional correspondence space accounts for 61.3% of the total variation in the Reagan corpus and somewhat less in the two aggregate corpora.[23] Yet, dimensionality in this context requires careful dissection and analysis before a coherent picture may be obtained.

In Figure 1, we can observe a cleavage in the discourse between domestic economic policy themes (federal budgetary reform, policy reform initiatives, and macroeconomic policy) and themes relating to human values, democracy, and national security. This is

19. While correspondence analysis is well established in the French literature (see (Benzecri 1973) and the journal *Cahiers de l'Analyse des Donnees*) its use has spread with the publication of English applications (Greenacre 1984; Greenacre and Underhill 1982; Weller and Romney 1990) and is occasionally used by political scientists (Blasius and Thiessen 2001) . Correspondence analysis using numerical data is available in several major statistical packages, including BMDP, SPSS, and SAS.

20. For this, correspondence analysis uses the "chi-squared distance," which resembles the Euclidean distance between points in physical space. (Here, chi-squared distance—which is distinct from the chi-squared statistic used to measure the significance of the words and tags—can be observed in Euclidean space by transforming the profiles before constructing the plots.) In correspondence analysis, each squared difference between coordinates is divided by the corresponding element of the average profile (where the profile is a set of frequencies divided by their total). The justification for using the chi-squared concept is that it allows one to transform the frequencies by dividing the square roots of the expected frequencies, thereby equalizing the variances. This can be compared to factor analysis, where data on different scales are standardized. For more detailed discussion and further geometric reasons for using the chi-squared distance in correspondence analysis, see Greenacre (1993, 34-36).

21. Correspondence analysis usually refers to the "inertia" of a table, which can also be called "association" (Weller and Romney 1990) . A corresponding chi-squared value can be obtained by multiplying the association value by the total n of the table.

22. The association and chi-squared statistic may be interpreted geometrically as the degree of dispersion of the set of rows and columns (or, profile points) around their average, where the points are weighted.

23. In total, four factors are identified in the correspondence analysis for Reagan (with the remaining factors obtaining percentage associations of 19.8, and 18.8, in ascending order). For Figure 2, five factors are identified, and the remaining percentage associations are 16.5, 15.1. and 12.3; for Figure 3, there are a total of six factors, with associations for the remaining factors of 15.1, 12.6, 10.4 and 7.1—again, in ascending order. (Usually, the dimensionality of the system is one less than the number of classes in the profile (Greenacre 1993, 14).)
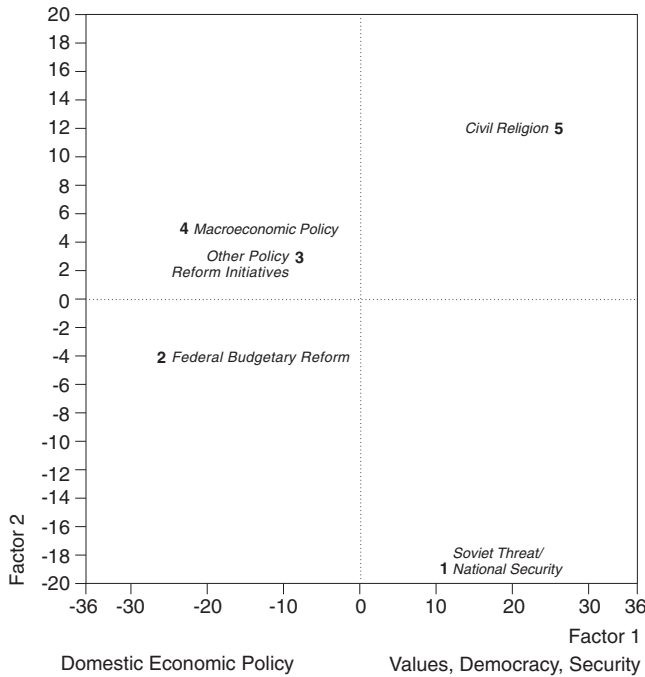
FIGURE 1. Correspondence Analysis of Ronald Reagan's State of the Union Speeches.

reflected on the horizontal axis. While we have not labeled it as such, the vertical axis might arguably reflect a second cleavage between domestic themes (in or near the upper quadrants) and foreign policy (in the lower quadrant).

Figures 2 and 3 are more informative and intriguing. Both figures exhibit a distinct grouping of themes relating to policy and policy process on one side of the horizontal axis. In the earlier period (Figure 2), these include the effect of the Great Depression on agriculture and taxes, relations between the executive and congressional branches, and policies on the nation's infrastructure. In the later period (Figure 3), these include relations and coordination between the executive and congressional branches, agriculture and business, macroeconomic policy (deficits, taxes, inflation, unemployment), and policies relating to education and welfare. The second pole of this first dimension, however, appears to reflect an historical shift in presidential discourse. Perhaps reflecting the experiences of the two world wars, the discourse for the earlier presidents focuses on democracy, values and peace (Wilson and Roosevelt's focus on democratic values, honoring the war dead, and international peace). Presidents of the post–World War II era certainly discuss foreign policy (especially in the context of the Cold War), but there is a greater motivational element to
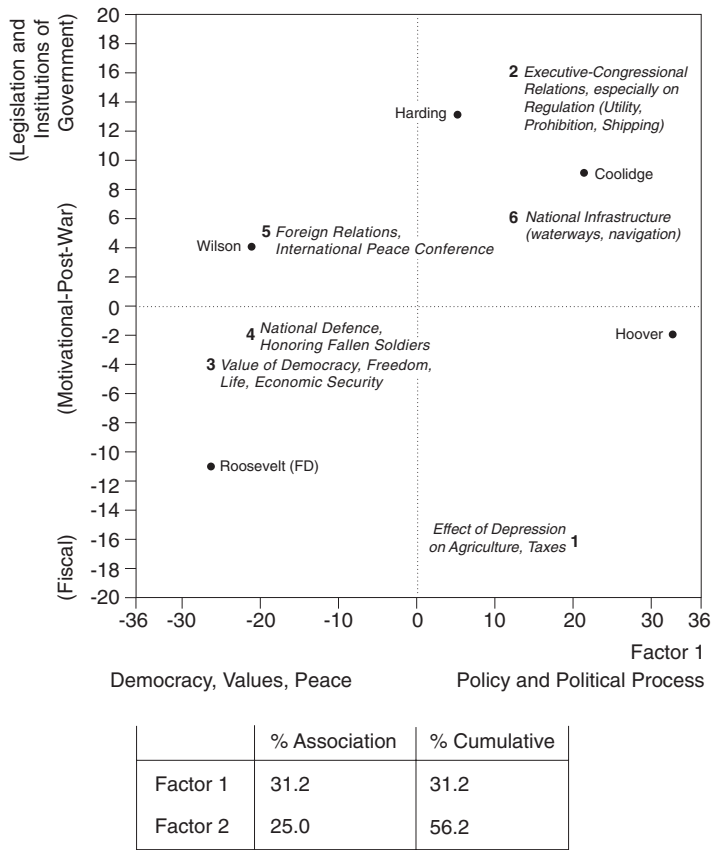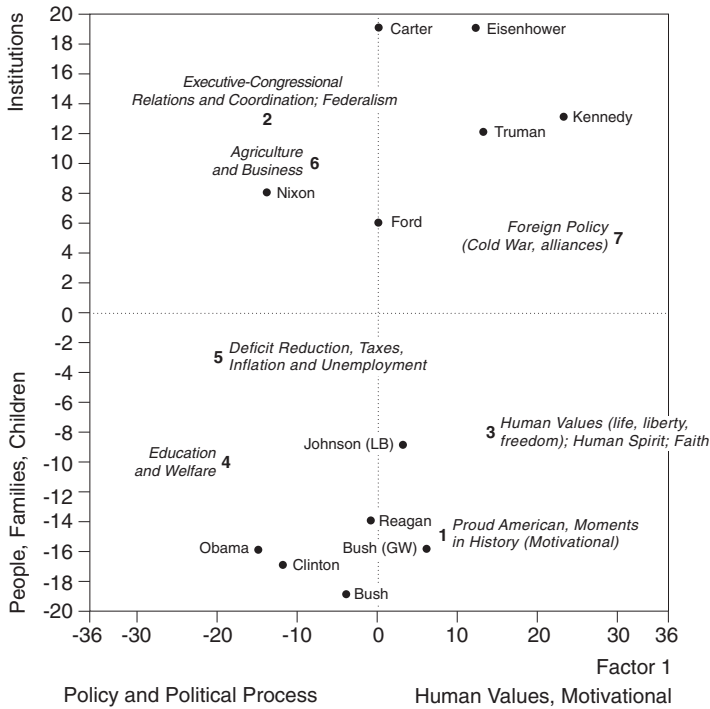
FIGURE 2. Correspondence Analysis of State of Union Speeches from W. Wilson to F. D. Roosevelt.

this rhetoric—as reflected in classes 1 and 3 (proud Americans, historical moments; human values of life and liberty, the human spirit, and faith).

The vertical dimensions for both Figures 2 and 3 tell a second part of the story. In the pre–World War II period, the "dimension" is less polar and more one of three distinct groups—as we indicate in the brackets: fiscal policy in the lower quadrant, postwar motivational rhetoric in the middle, and themes relating to legislation and institutions of government in the top quadrants. In the later period, we see two poles of the second dimension more clearly. In the top quadrants are thematic classes that broadly speaking focus on institutions (branches of government, business and agriculture, alliances in the Cold War). With the exclusion of class 5 (macroeconomic policy issues), classes 1, 3, and 4 all contain rhetoric pertaining to individuals, families, and children. The most intriguing element of this spatial representation of presidential rhetoric is that *all* presidents from Reagan onward are situated in the lower quadrants. That is, Reagan, both Bush presidents, Clinton, and Obama *all* have focused more of their rhetoric on issues relating to Americans as individuals, American families, and American children, and less on what

	% Association	% Cumulative
Factor 1	30.7	30.7
Factor 2	24.1	54.8

FIGURE 3. Correspondence Analysis of State of Union Speeches, H. Truman to B. Obama.

might be described as more institutions-based rhetoric. If we exclude Johnson (and bearing in mind his more northerly position), Reagan *appears* to be a discourse "pivot" between the two broad thematic groups.

What does this say about the distinctiveness of Reagan's rhetoric? While the findings are by no means definitive—particularly bearing in mind that our two-dimensional spatial representation does not capture the larger multidimensional space—they do appear to suggest that certain elements of Reagan's rhetoric resonated in the State of the Union addresses of subsequent presidents. While our earlier analysis suggests that Reagan's usage of "God" is quite distinctive, a more subtle shift toward discourse reflecting the importance of the individual, families, and children may indicate that he was a rhetorical pivot not so much in terms of civil religion rhetoric, but more in terms of shifting the discourse from the more abstract level of institutions to the more tangible level of personal experiences.[24]

24. One might extend this to include Jamieson's interpretation of "manly" and "womanly" speech, but this is beyond the scope of this paper (Jamieson 1988).

Discussion and Conclusion

By way of summarizing our findings, we return to three categories of our results typology—the classic, the unexpected, and the residue. Our "classic" result is that Reagan did indeed express civil religion rhetoric—which accords with anecdotal commentaries. Our approach, however, assigns an empirical weight to this rhetoric—for his seminal speeches and including speeches to conservative audiences, approximately 59% of the classified text focuses on civil religion. Even in the harder test of his State of the Union speeches (and controlling for the potentially vacuous "God bless" term), we still find that nearly half (48%) of Reagan's discourse fits within the rubric of civil religion. In comparison with the equivalent speeches of other modern presidents' Reagan is unparalleled both in his extensive use of civil religion rhetoric and his particular reliance on the term "God." Our "unexpected" results derive from our correspondence analysis of the State of the Union speeches of all the modern presidents. In a two-dimensional spatial map, the first dimension for both the pre– and post World War II presidents appears to divide policy and process from more emotive themes. For the earlier presidents, these abstract themes centered on democracy, values, and international peace (and thus reflect the war-time context). For the later presidents, these emotive themes contain what might be construed as motivational rhetoric (proud Americans, the human spirit, and so on). More intriguing is that the second (vertical) dimension for the post–World War II presidents illustrates that all presidents from Reagan onward focused more of their rhetoric on issues relating to Americans as individuals, as families, and as children—and less on institutional themes (branches of government, alliances in the Cold War). Arguably, Reagan may be a rhetorical pivot from which presidential rhetoric seems to have shifted from the more abstract focus on institutions to the more tangible focus on personal experiences. And, finally, we acknowledge the limitations of our findings by examining the "residue" or unexplained portion of our analysis: as seen from Tables 1, 3 and 4, our methodology is unable successfully to classify 17% of Reagan's State of the Union speeches, 29% of the State of the Union speeches of earlier presidents and 23% of the later presidents. Moreover, while the correspondence analysis reveals intriguing results, the two dimensions are ultimately limited by the inability to capture the spatial map of the higher dimensions. So, while our approach has accounted for a large part of the discourse in State of the Union addresses, there is more work to be done.

Taking on board these findings, how might our results contribute to our initial observations on the trend toward religious polarization between the Republican right and the Democratic left? Notably, our findings do not suggest that presidents of either party exhibit this religious polarization in their State of the Union addresses. From Reagan onward, presidential rhetoric of both parties appears to be converging, not diverging (as seen particularly in Figure 3). Whether the focus is more on policy and process, or on motivational discourse, recent presidents all tend to infuse their speeches with more references to personal experiences of individuals, families, and children, and fewer references to institutions. There is, of course, some degree of civil religion rhetoric as part of this discourse but less specific mention of "God" from post-Reagan presidents. To the

extent that this rhetoric by recent presidents is targeted at any common denominator, this might be construed as less of a religious nature and more one of personal motivation and inspiration.

References

Albanese, C. L. 1976. *Sons of the Fathers: The Civil Religion of the American Revolution*. Philadelphia: Temple University Press.

Auer, J. J. 1992. "Acting Like a President; or, What Has Ronald Reagan Done to Political Speaking?" In *Reagan and Public Discourse in America*, ed. M. Weiler and W. B. Pearce. Tuscaloosa: University of Alabama Press.

Bara, J., A. Weale, and A. Biquelet. 2007. "Analysing Parliamentary Debate with Computer Assistance." *Swiss Political Science Review* 13 (4): 577-605.

Bastin, G. 2002. Note sur la méthode Alceste. Mondes sociaux et mondes lexicaux [Note on the Alceste method. Social worlds and lexical worlds]. *Melissa*. http://www.melissa.ens-cachan.fr/spip.php?article200 (accessed April 13, 2012).

Bauer, M. 2000. "Classical Content Analysis: A Review." In *Qualitative Researching with Text, Image and Sound: A Practical Handbook*, ed. M. W. Bauer and G. Gaskell. London: Sage, 131-51.

Beaudouin, V., and S. Lahlou. 1993. "L'analyse lexicale, outil d'exploration des représentations. Réflexions illustrées par une quinzaine d'analyses de corpus d'origines très diverses" [Lexical analysis, a tool for exploring representations. Reflections illustrated by fifteen analyses of very diverse corpuses]. *Cahiers de Recherche de CREDOC*, no. 48 (September): 146.

Bellah, R. N. 1967. "Religion in America." *Daedalus* 96 (1): 1-21.

Bellah, R. N., and P. E. Hammond. 1980. *Varieties of Civil Religion*. San Francisco: Harper & Row.

Benzecri, J.-P. 1973. *L'analyse des données. Tome1: La Taxinomie. Tome 2: L'Analyse des Correspondances* [Data analysis. Volume 1: Taxonomy. Volume 2: Correspondance analysis]. Paris: Dunod.

Bicquelet, A. 2009. "On Referendums: A Comparison of French and English Parliamentary Debates Using Computer-assisted Textual Analysis." Ph.D. diss. Wivenhoe, University of Essex.

Bimes, T., and Q. Mulroy. 2004. "The Rise and Decline of Presidential Populism." *Studies in American Political Development* 18 (2): 136-59.

Blankenship, J., and J. K. Muir. 1992. "The Transformation of Actor to Scene: Some Strategic Grounds of the Reagan Legacy." In *Reagan and Public Discourse in America*, ed. M. Weiler and W. B. Pearce. Tuscaloosa: University of Alabama Press, 135-60.

Blasius, J., and V. Thiessen 2001. "Methodological Artifacts in Measures of Political Efficacy and Trust: A Multiple Correspondence Analysis." *Political Analysis* 9 (1): 1-20.

Blei, D. M., and J. D. Lafferty. 2006. "Dynamic Topic Models." Presented at the 23rd International Conference on Machine Learning, Pittsburgh, PA.

———. 2009. "Topic Models." In *Text Mining: Classification, Clustering, and Applications*, ed. A. Srivastava and M. Sahami. Boca Raton, FL: CRC Press, 71-94.

Bligh, M. C., J. C. Kohles, and J. R. Meindl. 2004. "Charting the Language of Leadership: A Methodological Investigation of President Bush and the Crisis of 9/11." *Journal of Applied Psychology* 89 (3): 562-74.

Brugidou, M. 1998. "Epitaphes, l'image de Francois Mitterrand à travers l'analyse d'une question ouverte posée à sa mort" [Epitaphs, Francois Mitterrand's Image: An Analysis of an Open Question Asked on His Death]. *Revue Française de Science Politique* 48 (1): 97-120.

———. 2003. "Argumentation and Values: An Analysis of Ordinary Political Competence Via An Open-Ended Question." *International Journal of Public Opinion Research* 15 (4): 413-30.

Brugidou, M., C. Escoffier, R. Le Dominique, H. Folch, S. Lahlou, P. Morin-Andreani, and G. Piat. 2000. "Les facteurs de choix et d'utilisation de logiciels d'analyse de données textuelles" [Parameters for chosing and using text mining software]. *Journées internationales d'analyse statistiques des données textuelles* JADT 2000: 373-80. http://lexicometrica.univ-paris3.fr/jadt/jadt2000/tocJADT2000.htm (accessed April 18, 2012).

Campbell, K. K., and K. H. Jamieson. 1985. "Inaugurating the Presidency." *Presidential Studies Quarterly* 15 (2): 394-411.

———. 1990. *Deeds Done in Words*. Chicago: University of Chicago Press.

———. 2008. *Presidents Creating the Presidency: Deeds Done in Words*. Chicago: University of Chicago Press.

Chidester, D. 1988. *Patterns of Power: Religion and Politics in American Culture*. Englewood Cliffs, NJ: Prentice-Hall.

Dunn, C. W., ed. 1984. *American Political Theology: Historical Perspective and Theoretical Analysis*. New York: Praeger.

Edel, W. 1987. *Defenders of the Faith: Religion and Politics from the Pilgrim Fathers to Ronald Reagan*. New York: Praeger.

Ellis, R. J., ed. 1998. *Speaking to the People: The Rhetorical Presidency in Historical Perspective*. Amherst: University of Massachusetts Press.

Foerstel-Branson, P. 2009. "The Study of Political Tradition Through Rhetoric: The Bush Doctrine, How it Relates to the September 11, 2001 Terrorist Attacks, and Its Comparison to Wilsonianism" Ph.D. diss., Claremont Graduate University.

Fowler, R. B., A. D. Hertzke, and Laura R. Olson. 1999. *Religion and Politics in America: Faith, Culture, and Strategic Choices*. Boulder, CO: Westview Press.

Gallup, G., and J. Castelli 1989. *The People's Religion: American Faith in the 90's*. New York: Macmillan and Collier Macmillan.

Greeley, A. M. 1989. *Religious Change in America*. Cambridge, MA: Harvard University Press.

Greenacre, M., and T. Hastie 1987. "The Geometric Interpretation of Correspondence Analysis." *Journal of the American Statistical Association* 82 (398): 437-47.

Greenacre, M. J. 1984. *Theory and Applications of Correspondence Analysis*. London: Academic Press.

———. 1993. *Correspondence Analysis in Practice*. London: Academic Press.

Greenacre, M. J., and L. G. Underhill 1982. "Scaling a Data Matrix in Low-Dimensional Euclidean Space." In *Topics in Applied Multivariate Analysis*, ed. D. M. Hawkins. Cambridge, UK: Cambridge University Press, 183-266.

Guerin-Pace, F. 1998. "Textual Analysis, An Exploratory Tool for the Social Sciences." *Population: An English Selection, special issue of New Methodological Approaches in the Social Sciences* 10 (1): 73-95.

Hughey, M. W. 1983. *Civil Religion and Moral Order: Theoretical and Historical Dimensions*. Westport, CT: Greenwood Press.

Jamieson, K. 1988. *Eloquence in an Electronic age: The Transformation of Political Speechmaking*. New York: Oxford University Press.

Jasinski, J. 1992. "Antithesis and Oxymoron: Ronald Reagan's Figurative Rhetorical Structure." In *Reagan and Public Discourse*, ed. M. Weiler and W. B. Pearce. Tuscaloosa: University of Alabama Press.

Jenny, J. 1997. "Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. Etat des lieux et essai de classification" [Methods and formalized practices for content and discourse analysis in contemporary French sociological research. State of the art and proposal for a taxonomy]. *Bulletin de Méthodologie Sociologique (B.M.S.)* 54 (March): 64-112.

Kronberger, N., and W. Wagner 2000. "Keywords in Context: Statistical Analysis of Text Features." In *Qualitative Researching with Text, Image and Sound: A Practical Handbook*, ed. M. W. Bauer and G. Gaskell. London: Sage, 299-317.

Lahlou, L. 1996. "A Method to Extract Social Representations from Linguistic Corpora." *Japanese Journal of Experimental Social Psychology* 35 (3): 278-91.

Lahlou, S. 1995a. "Penser Manger. Les représentations sociales de l'alimentation. Paris" [Food for thought: Social representations of eating]. Ph.D. diss., École des Hautes Études en Sciences Sociales.

Lahlou, S. 1995b. "Vers une théorie de l'interprétation en analyse des données textuelles" [Towards a theory of interpretation in text mining]. In *3rd International Conference on Statistical Analysis of Textual Data*, ed. S. Bolasco, L. Lebart, and A. Salem. Rome: CISU.

Lebart, L., and A. Salem. 1994. *Statistique Textuelle* [Statistics for text mining]. Paris: Dunod.

Lejon, K. O. U. 1988. *Reagan, Religion and Politics*. Skillingaryd, Sweden: Lund University Press.

Lim, E. T. 2002. "Five Trends in Presidential Rhetoric: An Analysis of Rhetoric from George Washington to Bill Clinton." *Presidential Studies Quarterly* 32 (2): 328-48.

———. 2008. *The Anti-Intellectual Presidency: The Decline of Presidential Rhetoric from George Washington to George W. Bush*. Oxford, UK: Oxford University Press.

Nickerson, R. S. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises." *Review of General Psychology* 2 (2): 175-200.

Noel-Jorand, M. C., M. Reinert, M. Bonnon, and P. Therme. 1995. "Discourse Analysis and Psychological Adaptation to High Altitude Hypoxia." *Stress Medicine* 11 (1): 27-39.

Noel-Jorand, M. C., M. Reinert, S. Giudicelli, and D. Dassa. 1997. "A New Approach to Discourse Analysis in Psychiatry, Applied to Schizophrenic Patient Speech." *Schizophrenia Research* 25 (3): 183-98.

———. 2004. "Schizophrenia: The Quest for a Minimum Sense of Identity to Ward Off Delusional Disorder." *Canadian Journal of Psychiatry* 49 (6): 394-98.

Pierard, R. V., and R. D. Linder. 1988. *Civil Religion and the Presidency*. Grand Rapids, MI: Zondervan.

Popping, R. 2004. E-mail correspondence with authors, October 29.

Putnam, R. D., and D. E. Campbell. 2010. *American Grace: How Religion Divides and Unites Us*. New York: Simon & Schuster.

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54 (1): 209-28.

Reinert, M. 1983. "Une methode de classification descendante hierarchique: application a l'analyse lexicale par contexte" [A method of descending hierarchical classification: Application to the analysis by lexical context]. *Les Cahiers de l'Analyse des Donnees* 8 (2): 187-98.

———. 1987. "Classification descendante hiérarchique et analyse lexicale par contexte: application au corpus des poésies d'Arthur Rimbaud" [Descending hierarchical classification and context-based lexical analysis: Application to the corpus of poems by A. Rimbaud]. *Bulletin de Méthodologie Sociologique* 13 (January): 53-90.

———. 1990. "ALCESTE. Une methodologie d'analyse des donnees textuelles et une application: Aurelia de Gerard de Nerval" [ALCESTE: A methodology for analyzing textual data and application: Aurelia de Gerard de Nerval]. *Bulletin de Methodologie Sociologique* 26 (1): 24-54.

———. 1993. "Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars" ["Lexical Worlds" and their "logics" through the statistical analysis of a corpus of narratives of nightmares]. *Langage et Société* 66: 5-39.

———. 1998a. ALCESTE users' manuel (English version). Toulouse, France: Image.

———. 1998b. "Quel objet pour une analyse statistique du discours ? Quelques réflexions à propos de la réponse Alceste" [What is the object of a statistical analysis of discourse? Some reflections about the Alceste solution]. In *Proceedings of the 4th JADT (Journées d'Analyse des Données Textuelles)*, ed. S. Mellet. Nice, France: Université de Nice, 557-70.

———. 2003. "Le rôle de la répétition dans la représentation du sens et son approche statistique dans la méthode Alceste" [The function of repetition in the representation of meaning and its statistical approach in the Alceste method]. *Semiotica* 147 (1-4): 389-420.

Ritter, K., and D. Henry. 1992. *Ronald Reagan: The Great Communicator*. New York: Greenwood Press.

Schonhardt-Bailey, C. 2005. "Measuring Ideas More Effectively: An Analysis of Bush and Kerry's National Security Speeches." *PS: Political Science and Politics* 38 (3): 701-11.

———. 2006. *From the Corn Laws to Free Trade: Interests, Ideas and Institutions in Historical Perspective*. Cambridge, MA: MIT Press.

Skinner, Q. 1974. "Some Problems in the Analysis of Political Thought and Action." *Political Theory* 2 (3): 277-303.

Toolin, C. 1983. "American Civil Religion from 1789 to 1981: A Content Analysis of Presidential Inaugural Addresses." *Review of Religious Research* 25 (1): 39-48.

Tulis, J. K. 1987. *The Rhetorical President*. Princeton, NJ: Princeton University Press.

U.S. Congress, Joint Committee on Printing. 2005. *Ronald Reagan: Late a President of the United States, Memorial Tributes Delivered in Congress*. Washington, DC: Government Printing Office.

Wald, K. D., and A. Calhoun-Brown. 2007. *Religion and Politics in the United States*, 5th ed. Lanham, MD: Rowman & Littlefield.

Weale, A. 2010. "Political Theory and Practical Public Reasoning." *Political Studies* 58 (2): 266-81.

Weiler, M., and W. B. Pearce. 1992. "Ceremonial Discourse: The Rhetorical Ecology of the Reagan Administration." In *Reagan and Public Discourse in America*, ed. M. Weiler and W. B. Pearce. Tuscaloosa: University of Alabama Press, 11-42.

Weller, S. C., and A. K. Romney. 1990. *Metric Scaling: Correspondence Analysis*. London: Sage.

Wilcox, C., and C. Robinson. 2007. "Prayers, Parties, and Preachers: The Evolving Nature of Political and Religious Mobilization." *From Pews to Polling Places: Faith and Politics in the American Religious Mosaic*, ed. J. M. Wilson. Washington, DC: Georgetown University Press, 1-28.

Wolfe, A. 2010. "Political Science, Democray, and Religion." In *Religion and Democracy in the United States: Danger or Opportunity?*, ed. A. Wolfe and I. Katznelson. New York: Russell Sage Foundation and Princeton University Press, 19-45.

Appendix: The Alceste Methodology

Overview

Alceste was developed by Max Reinert (Reinert 1983, 1998, 2003) and has been applied in sociology, psychology, and political science (Bara, Weale, and Biquelet 2007; Bauer 2000; Brugidou 1998, 2003; Guerin-Pace 1998; Lahlou 1996; Noel-Jorand, Reinert, Bonnon et al. 1995; Noel-Jorand, Reinert, Giudicelli et al. 1997, 2004; Schonhardt-Bailey 2005, 2006). Its specific advantages have been discussed elsewhere (Jenny 1997; Brugidou et al. 2000).[25] Alceste does not require any precoding but its application is constrained in that it cannot analyze very large corpora.[26] In contrast to applications of topic modeling where, for instance, vast libraries of documents with diverse topics might be analyzed, Alceste is suited to more focused research projects—for example, documents relating to a particular area of policy (monetary policy, trade policy, health) or the speeches of politicians that may be expected to contain broadly similar themes (State of Union addresses). Moreover, whereas topic modeling requires very little substantive knowledge of the topics themselves, Alceste is most effective when it is joined with expert substantive knowledge of the subject matter, since contextual knowledge is often essential for interpreting the form of argumentation as well for extending the analysis into its more specialist usages (e.g., the *Tri-Croisé*).

Alceste was initially designed to measure what Max Reinert calls the "lexical worlds." As Reinert explains, "we assume that the speaker, during his speech, is investing successive different worlds and these worlds, by imposing their properties, thereby impose a specific vocabulary. Therefore, the statistical study of the distribution of this vocabulary should be able to trace these 'mental rooms' that the speaker has successively inhabited; traces perceptible in terms of 'lexical worlds' " (1987). In other words, a lexical world is a specific vocabulary, which inherits its properties from what the subject is

---

25. See also http://www.cmh.pro.ens.fr/bms/arcati/BMS54-Jenny-New.htm (accessed April 13, 2012) for a list of the web sites and papers comparing text mining software.

26. Although subsequent versions may allow a larger corpus, Alceste 4.7 requires that the corpus not exceed 15 mb.

talking about—for example, if the text is about medicine, there will be many medical terms. Conversely, if there are many medical terms, this is a cue that the text may be about medicine.

By purely distributional means, the sets of words that go together in the discourse are isolated and represented to the researcher as a trace of some "lexical world" that remains to be interpreted. Hence, the basic idea of the software is to find "lexical worlds" in the speaker's discourse. The software accomplishes this using only a statistical approach to analyze the distribution of words in the corpus, while remaining completely deaf to the meaning of words themselves. The only semantic aspects inbuilt in the software are some grammatical dictionaries that enable reducing verbal forms to a single root (reducing all the flexions of a single verb to its radical, or names in plural to singular) and classifying words into various grammatical classes (nouns, verbs, articles etc.) so as to eliminate function words (articles, some prepositions) in the analysis.

So, as presidents run through semantic fields when producing discourse, their statements provide us with a lexical distribution that reflects the content of these fields (their lexical worlds). By classifying together the statements that contain similar words, we can hope to understand what semantic territories were behind the construction of the observed discourse.

Alceste operationalizes these notions of "statements," "words," and "similarity." Statements are approximated by ECUs, which are natural sentences or natural fragments of sentences delimited by punctuation so as to have similar length. Alceste constructs a dictionary of "lexical forms" ("lexemes") that are lemmatized words, more useful for our purpose in terms of semantics.

To assess similarity between statements, Alceste constructs a matrix that crosses ECUs and lexemes, where the cells sign the presence or absence of that lexeme in the ECU. Alceste then operates on this matrix a descending classification, which produces classes of similar context units. The descending classification technique used maximizes the similarity between statements in the same class *and* also maximizes the difference between the classes.

In the end, the analyst is provided with a series of classes and of statistical cues in the form of typical words, typical ECUs, typical authors, and so on. This provides basis for "interpreting" the classes as lexical worlds.

Operations in Alceste are statistical, transparent, and reproducible, until the final moment of interpretation, where the analyst assigns a label to each set of specific vocabulary, which was identified as a lexical world by the software, on the basis of co-occurrences and distribution patterns.

The Principle of Descending Classification

The core of Alceste is based on descending classification of text segments. The objective is to sort the ECUs in a partition of classes that are each homogeneous and as different as possible from one another. The initial table consists of as many lines as ECUs and as many columns as lexemes chosen for analysis. At the intersection of row i and

column j, the value is either 1 if ECU_i contains at least one occurrence of the $lexeme_j$; 0 otherwise.

The classification is a recursive algorithm. The first class comprises the total set of context units. The program then attempts to partition that class into two further classes that are each as homogeneous as possible and as different as possible from one another. This overlap between classes can be measured by the χ^2 of a table with two rows (one for each class) and as many columns as there are lexemes. A cell at the intersection of row i and column j will contain the number k_{ij} of context units of class I containing lexeme j.

$$\text{Lexeme } j$$

	...	K_{2j}	...		K_2
Class 2					
Class 3		K_{3j}	...		K_3
	...				

$$K_j$$
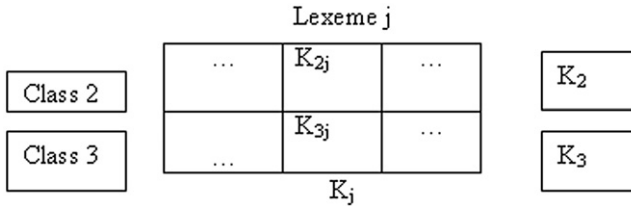
With, for example:

$$k_2 j = \sum_{i \in I_2} k_{ij}; \; k_2 = \sum_{i \in I_1} k_{2j}; \; k_j = k_{2j} + k_{3j};$$

The χ^2 can written as:

$$chi2 = k_2.k_3 \sum_{j \in J} \left(k_{2j}/k_2 - k_{3j}/k_3 \right)^2 \Big/ k_j$$

The objective is to search, among all possible partitions in two classes, for the one that maximizes the χ^2 of this table. In practice, the algorithm

1. Calculates the first factor of the Correspondence Factor Analysis of the table (R_j space, with the χ^2 metrics (Benzecri 1973);
2. Slides the orthogonal hyperplane along that axis until it reaches a position that maximizes the inter-class inertia of the two parts of the cloud of context units that are separated by this hyperplane;
3. Optimizes the partition with a local exchange algorithm by swapping context units individually around the hyperplane.

Following an iterative process, the descending hierarchical classification method decomposes the classes until a predetermined number of iterations fail to result in further significant divisions. The result is a hierarchy of classes that may be schematized as a tree diagram.

One advantage in using of descending classification (e.g., over ascending clustering techniques) is robustness. Textual data matrices are "scarce matrices" (mostly constituted of zeros, since only a small part of the whole vocabulary is used in each sentence) and, as such, are very sensitive to artifacts. A few very similar sentences (e.g., using a fixed expression) will create a strong correlation that can "pull" an artifactual class. In the

descending classification algorithm used, such effects stay local and do not propagate to the whole analysis.

The Four Steps of the Method

Alceste has been described as a "methodology" insofar as it "integrates a multitude of highly sophisticated statistical methods," (Kronberger and Wagner 2000, 306). The detail of the algorithms has been extensively described (Beaudouin and Lahlou 1993; Lebart and Salem 1994; Lahlou 1995a; Reinert 1983, 1987, 1990, 1993, 1998a, 1998b, 2003).

Step A: The dictionaries of the vocabulary are created. It identifies word categories (verbs, nouns, etc.) that are useful because only some categories are being used in analysis (excluding "tool words" such as articles or pronouns). Vocabulary is parsed using a grammatical dictionary and reduced into a dictionary of (lemmatized) lexical forms. All forms of a verb are converted to their infinitive, plurals are reduced to singular, and some variants of the same lexical root are reduced to the root. These forms are the basic "lexemes" upon which calculations are made.

Couples (two consecutive forms) and repeated segments (consecutive series of forms) are also found and listed in this step. They are useful in illustrating the classes at a later step. Calculations of frequencies of the forms are also done at this stage, which may be used in further steps (e.g., for eliminating from the analysis words that occur only once).

Step B: The software cuts the text into "statements." These are obtained empirically as ECUs, in practice sentences or parts of sentences cut by natural punctuation. There are three types of context units. ICUs are the pieces of text that constitute the corpus to be analyzed (here, the speeches of presidents). The ICU is essentially the sampling unit— that is, a preexisting division of the text and is specified by the user.

ECUs are the atomic pieces of text. The ECU is a "gauged sentence," which the program automatically constructs based upon and punctuation in the text.[27] An ECU is delimited by a punctuation sign and contains at least 15 occurrences of words.

In the course of segmentation, the software will also use somewhat larger Context Units (CUs), made by concatenation of several succeeding ECU.

In text analysis, a persistent and difficult issue concerns the optimal length of a "statement" as a semantic unit. Various possible solutions for a contextual unit would be a sentence, a paragraph, a piece of sentence, and so on. Alceste resolves this issue by not trying to identify directly the statement length; rather it produces classifications that are *independent* of the length of the statements.

For this, Alceste creates two classifications, each using units of different lengths of CUs, and then retains only the classes that appear in both classifications: these classes are independent of the length of statements. In practice, CUs in each classification are constructed as concatenations of ECUs that have a minimal length of $N1$ *active words* (e.g., 12), and in the second classification a length of $N2$ active words (e.g., 18). An active word

---

27. Popping notes that the ECU is akin to the "recording unit" used in other programs, where it is usually defined by the researcher (2004).

is a lexical root or "lexeme" (see Step A) of a noun, verb, or adverb, which occurs at least two times in the corpus. Alceste compares the two classifications, on the basis of ECUs in the classes. Only the classes that appear in both classifications are retained for analysis. The resulting classification is stable and, as said, independent of the length of the CUs. This leaves a number of ECUs unclassified, thereby approximating a measure of goodness-of-fit.

The stability of the partitioning is measured by constructing a table that crosses all the classes (including all levels of nodes) obtained in the first classification and all the classes obtained in the second classification. In each cell (Cpq) of this table is the χ^2 associating the two classes p and q. The result is a "signed chi-square table"—that is, a data table with the positive and negative links between the classes. This table helps to select the classes that share the higher number of ECUs. Interestingly, this table not only enables the program to retain the stable partitions, but also to obtain an empirical solution as to where to stop descending classification in the classification tree: when stems are not stable, the tree truncates at the higher node.

This description, although limited, provides an indication of the elegance of the algorithm and the statistical underpinnings necessary to cope with the specific type of matrices encountered in textual data analysis ("scarce matrices"). This accounts for this specific software's exceptional robustness. For a detailed exposition of the algorithm, see (Bastin 2002; Lahlou 1995a; Reinert 1983); for a step-by-step explanation of each stage of the analysis, see (Reinert 1990), and for a simple illustration (Kronberger and Wagner 2000).

Steps C and D provide various auxiliary calculations to assist in the interpretation and description of classes. Most suggestive are the lists of the most representative vocabulary of each class (lexemes ordered by decreasing χ^2), and selected ECUs that best represent the class, based on their lexical content.

A correspondence analysis is also performed in this step to provide a comprehensive representation of the semantic field that situates the relative positions of the classes and lexemes.