What We Can Learn about Political Speech from the Pulpit from a Large Corpus of Sermons

Steven Morgan

August 30, 2018

Abstract

This document illustrates the representativeness of the sermon corpus data, as well as lays out potential research questions the data can be used to address. Due to memory issues, the data analyzed (briefly) herein only include sermons preached from 2011-2018. This results in a corpus of 73,969 sermons composed of over 100 million words from 41 different denominations (all of which are Christian). The sermons were uploaded by 2,420 members of the clergy. The full corpus spans 2000-2018 and comprises just over 160,000 sermons. This dataset allows the unique opportunity to test theories of political communication from the pulpit without relying on surveys of pastors of clergy. However, this corpus is certainly not representative, as sermons delivered by evangelical Protestant pastors (especially Baptists) are over-present while homilies delivered by Catholic priests are under-present in the corpus.

Introduction

All sermons were scraped from sermoncentral.com, a repository for members of the clergy to upload their sermons. Importantly, each sermon is accompanied by information on the name of the pastor, the denomination of the pastor, and the date the sermon was preached. This information was scraped as well. The website also lists the profiles of all contributors. This includes a picture of the pastor, the name of the pastor's church, and in some instances the location of the church. This document proceeds with two sections. The first section presents descriptive statistics for the corpus. The purpose of this sermon is to assess the representativeness of the corpus at a high-level. I am particularly concerned with the distribution of sermons uploaded by pastors, by year, and by denomination. The second section identifies research questions that can be tested, as well as initial forays into how a research design addressing these questions may come to fruition.

Descriptives of Sermon Corpus

Due to memory issues, initial analysis was limited to a subset of the corpus. While the full corpus is composed of over 160,000 sermons, I present initial analysis based on all sermons preached from 2011-2018 (73,969 sermons). This is not a representative sample of the full corpus, but issues with dealing with large quantities of text necessitated this subset for this initial dive. Note that I will refer to this subset as "the corpus" throughout. The corpus is comprised of just over 100 million words, with a vocabulary (number of unique words) of about 4.5 million. Figure 1 presents the distribution of word counts in each sermon. Figure 2 presents the distribution of unique word counts in each sermon. Both are relatively Gaussian. This is somewhat intuitive: even though this is considered a specialized corpus, at the lexical level words spoken should overlap frequently (even across denominations). And, true to text-as-data, there exists a long right tail, since many corpora exhibit some power law tendencies. This would be especially true at the unigram level (a few words appear very often, some words appear moderately frequently, and most words are rare in the corpus).

Moving away from a lexical descriptive approach, it is important to understand the distribution of pastors in the corpus. Ideally, there would exist sufficient variation among who is uploading these sermons. Ex ante, this was my greatest concern. Deciding to upload a corpus to a website/database may result in clustering of sermons by pastor, since this is not a direct window into

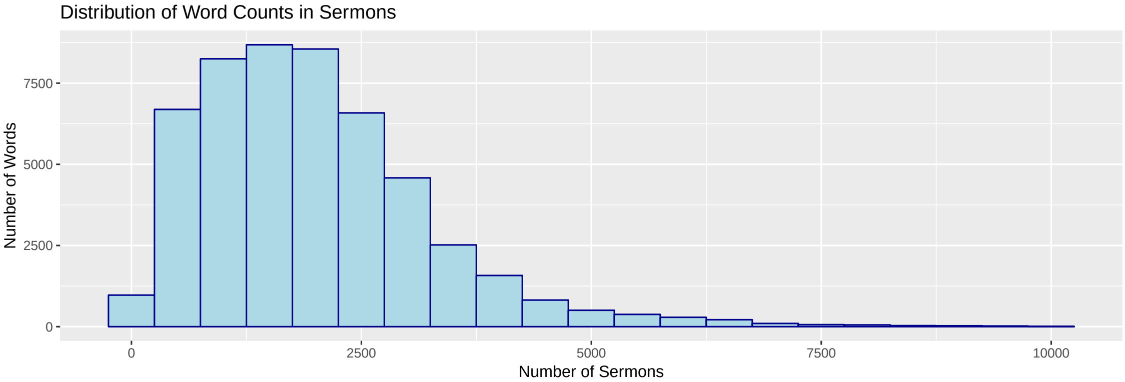

Figure 1: The distribution of words in sermons is somewhat normal; however, there is definitely a right-ward skew. Few sermons are less than 500 words, and the majority are between 750 and 2,000 words. A minority of sermons are over 5,000 words.

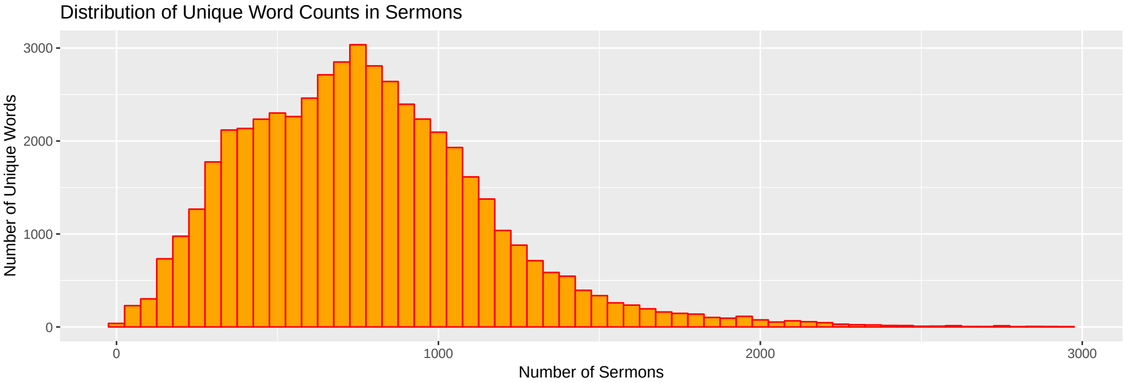

Figure 2: The distribution of unique words in sermons is similar to the distribution of total words across sermons. It is relatively normal, and there exists a somewhat long right tail. Few sermons are less than 500 words, and the majority are between 750 and 2,000 words.

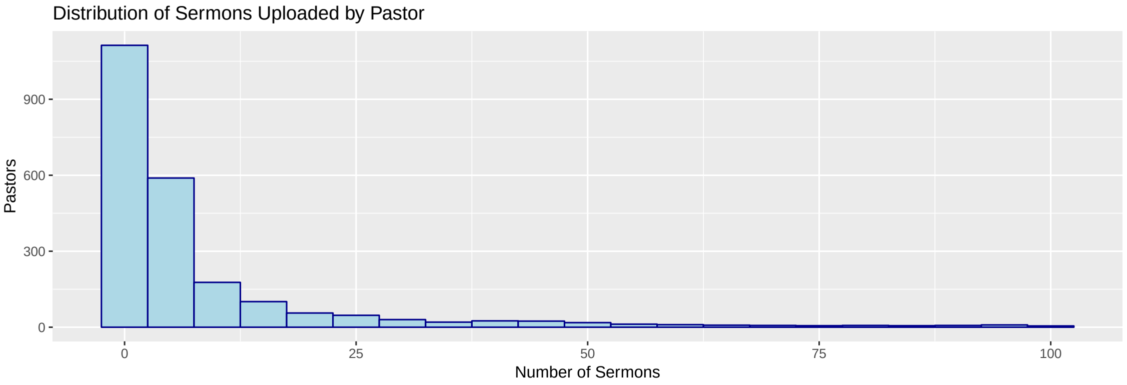

Figure 3: Most pastors upload fewer than 10 sermons. There exist some outliers, publishing upwards of 100 sermons.

sermons for the entire religious landscape. However, this issue does not seem as bad as initially thought. Figure 3 presents the distribution of the number of sermons uploaded by individual pastors. The distribution approximates a Poisson distribution, which makes sense since each sermon can be thought of as a count. The variation in the number of pastors uploading sermons (over 2,400) should allow for reasonable inference about the effects of denomination on political speech, without a great deal of concern that a few pastors are potentially dominating results.

Representativeness across denominations is desirable as well. This corpus can give leverage on answering questions on political speech across a variety of denominations if the breakdown of sermons by denomination reasonably approximates the US religious landscape. Table 1 presents the distribution of sermons by denomination, presenting both the frequency of sermon by denomination as well as the proportion of the corpus. Initially, Baptists are overrepresented compared to the general population (30.3% of the corpus while about 16.0% of Christians affiliate with Baptists) based on the 2016 Pew Religious Landscape Study. Catholics are underrepresented (only 1.2% of the corpus compared to about 29.5% of the US Christian population). This makes sense, as the hierarchical nature of the Catholic Church as well as higher levels of education usually attained by Catholic Priests should dissuade Catholic priests from making sermons publicly available. Similarly, Catholic churches rarely broadcast mass (outside of the Vatican), while this practice is common across multiple media platforms for evangelical Protestant churches.

Since this corpus is comprised of documents over a nontrivial period of time, it is important to understand how the data may change over time. Table 2 presents the number of sermons uploaded by year. The table illustrates how remarkably stable the upload stream was for 2011-2018. Note that data was no longer collected after June 30, 2018, thus the smaller amount for 2018. There appears to be a slight decrease in the number of sermons in the corpus in 2016 and 2017 from previous years. However, it is difficult to say if the content of what is uploaded is sufficiently different by year at this level of analysis.

Research Questions

Studies on political speech from the pulpit have nearly all drawn on surveys of pastors or members of their congregations. This is problematic for two reasons. First, the prohibitive costs of running a survey has forced researchers to focus on one or two denominations for a study. We know a lot about Southern Baptist pastor attitudes, but not very much about most Protestant denominations. Second, all survey items tapping political speech rely on recall of whether or not one spoke out politically and willingness to divulge this information to a stranger. The corpus allows for an unobtrusive purview into what political content members of the clergy are gifting their members. Additionally, the size of the corpus allows for the analysis of how political speech varies by denomination. The literature tells us that mainline Protestants speak out more politically due to their attention paid to social justice issues. However, these conclusions are all based on survey data. It may be the case that mainline Protestant pastors are more likely to state that they

Table 1: Breakdown of sermons across Christian denominations.

Denomination	# of Sermons	% of Corpus
Adventist	30	0.06
Anglican	1,012	1.99
Apostolic	148	0.29
Assembly Of God	2,198	4.31
Baptist	15,436	30.29
Bible Church	300	0.59
Brethren	271	0.53
Calvary Chapel	765	1.50
Catholic	621	1.22
Charismatic	626	1.23
Christian Church	914	1.79
Christian Missionary Alliance	562	1.10
Christian/Church Of Christ	3,973	7.80
Church Of God	776	1.52
Congregational	147	0.29
Disciples Of Christ	32	0.06
Episcopal	149	0.29
Episcopal/Anglican	38	0.07
Evangelical Free	591	1.16
Evangelical/Non-Denominational	5,019	9.85
Foursquare	600	1.18
Free Methodist	59	0.12
Friends	180	0.35
Holiness	879	1.72
Independent/Bible	3,173	6.23
Lutheran	1,017	2
Mennonite	15	0.03
Methodist	655	1.29
Nazarene	1,103	2.16
Not listed	308	0.60
Orthodox	3	0.01
Other	1,686	3.31
Pentecostal	3,455	6.78
Presbyterian/Reformed	1,675	3.29
Salvation Army	305	0.60
Seventh-Day Adventist	57	0.11
United Methodist	1,683	3.30
Vineyard	52	0.10
Wesleyan	453	0.89

Table 2: Number of Sermons in the Corpus by Year

Year	# of Sermons	% of Corpus
2011	7,514	14.74
2012	6,926	13.59
2013	7,594	14.90
2014	8,309	16.30
2015	7,161	14.05
2016	5,470	10.73
2017	4,605	9.04
2018	3,389	6.65

speak out politically to their congregations because they expect that they should. Especially juxtaposed against evangelical Protestants' connection to political conservatism and the GOP, I would not be surprised to find that we have discounted how much evangelical pastors speak out politically.

Political speech is a broad topic. I am interested in specific policy/issue frames pastors portray. The size of the corpus coupled with the fact that this is text data allows me, for example, to analyze not just whether or not pastors talk about abortion but how they frame it. There is growing literature on "rights talk", the idea that the Christian Right has taken up liberals' strategy of winning political battles on the basis of protections of individual rights (Lewis 2016). This is in stark contrast to evangelicals' focus up until the early 1980s on safeguarding morality and protecting the community. Conservatives have also taken up this cause. Hollis-Brusky gives great deal to the rise of the Federalist Society and a revamping of rights talk among conservatives as a legal movement (2015). Decker goes as far as calling the conservative legal movement witnessed over the last three decades a "revolution" and stating that it effectively has "remade government" starting with the Reagan years.

It would be interesting to analyze what role religious elites had in ushering in the "rights era" of conservatism. Lewis gives a very detailed account of the role of Christian legal advocacy groups and denominations. It is less clear the mechanism that allowed for elites to set out a strategy that did effectively trickle down to the masses. Members of the clergy may have played a role in communicating conservative trends from denomination officials to the masses.

This corpus may serve as a useful baseline corpus to measure religious speech. "Dog-whistling" is very tough to measure, and all of the work on religious coded language relies on experiments or qualitative accounts (Turek 2014). However, it may be possible to measure just how much religious language is employed by politicians in their speeches over time. For example, in Bush's 2002 State of the Union, the president stated: "there's power, wonder-working power, in the goodness and idealism and faith of the American people." While this may seem like an awkward phrase to the majority of the American people, the first half is actually a refrain from the popular evangelical hymn "There is Power in the Blood." A typical dictionary-based approach to religious language would miss these multivocal religious communications, because by their very nature they are tough to identify (and even more difficult to identify a priori). A baseline corpus coupled possibly with machine learning approaches (either active learning or the use of word embeddings) may be able to tease out these subtle but important communications by political elites.