Laura B. Bell
February 10, 2019
Corpus Linguistics Fall 2013: Final Project Summary

Introduction

For Computational Corpus Linguistics, I completed a project using Python that analyzed the C-Span State of the Union Address Corpus, available through NLTK. The project looked at tf-idf within the SOTUs over time, and by party

Figures 1 and 2 provide the results of this analysis.

Top 20 Terms by Decade							
	1940s	**1950s**	**1960s**	**1970s**	**1980s**	**1990s**	**2000s**
#1	1947	1953	vietnam	\xa1	ve	tonight	.
#2	reconversion	shall	tonight	\xa1\xa6	tonight	ought	applause
#3	liquidation	1954	1964	xand	1982	ve	terrorists
#4	1946	imperialism	nam	xa	afghanistan	21st	iraq
#5	receipts	survivors	viet	\xa1\xa7	nicaragua	bosnia	afghanistan
#6	1945	communist	1966	1974	1980	parents	iraqi
#7	demobilization	suffrage	1965	1973	m	reform	saddam
#8	authorizations	1955	1961	\xa1@	re	kids	tonight
#9	wartime	pg	1968	1975	hollings	thank	11th
#10	estimated	communists	1967	xthe	ll	toughest	al
#11	expenditures	constantly	1962	1976	sandinistas	lady	hussein
#12	considerable	subversion	negroes	92d	1986	saddam	qaeda
#13	navy	armaments	disarmament	xnot	pages	got	terrorist
#14	lease	adequate	shall	1971	1983	applause	regime
#15	occupation	atomic	color	crude	soviets	re	coalition

#16	shipping	highway	countryside	seventies	isn	class	prescription
#17	--	preparation	vietnamese	tonight	"	covenant	ve
#18	vj	statehood	communist	ve	43	11	thank
#19	appropriations	mobilization	selma	sixties	1984	iraq	2006
#20	adequate	postal	communists	1977	deterrence	internet	seniors

Figure 1: Top terms by decade

Top 20 Terms by Party		
	Republican	**Democratic**
#1	.	ought
#2	\xa1	reconversion
#3	\xa1\xa6	receipts
#4	regime	liquidation
#5	al	internet
#6	xand	demobilization
#7	homeland	rent
#8	alternative	brady
#9	coalition	millennium
#10	ballistic	1964
#11	involves	bosnia
#12	1975	nam
#13	11th	lobbyists
#14	qaeda	financed
#15	iraqis	viet
#16	younger	gore

#17	"	1966
#18	2001	scholarships
#19	xa	empowerment
#20	1982	1968

Figure 2: Top terms by party

Background

Before drawing any conclusions from these results, some basic information about the corpus would be helpful.

The majority of the corpus consists of one speech per year between 1945 and 2006, but there are a few exceptions: there is no transcript for 1952, and four years (1963, 1965, 1991, and 2001) are given double representation. In each of the years with double representation, one speech is the normal SOTU and the other is a special SOTU given after an event of considerable significance. These events are: 1) 1963: The assassination of President Kennedy; 2) 1965: The civil rights marches that resulted in “Bloody Sunday”; 3) 1991: The success of Operation Desert Storm; and 4) 2001: The terrorist attacks of September 11th.

Interesting results and possible explanations

statehood: 'Statehood' clocks in at No. 18 in the 1950s, likely because Hawaii and Alaska both became states in 1959.

1960s: This whole batch is impressively transparent. These results are practically a SparkNotes cheatsheet of 1960s events: 1) the Vietnam war ('vietnam', 'viet', 'nam', 'vietnamese', 'communist', 'communists'); 2) the civil rights movement ('negroes', 'selma', possibly 'color'); and, lamely, 3) the years occuring in this period ('1961', '1962', '1964', '1965', '1966', '1967', '1968'.)

1970s: Meanwhile, this batch appears to be tainted by a problem with the encoding that allowed source code to appear on the surface (e.g. '\u0105'). This also allowed words like 'america', 'nation', and 'world' -- words that would definitely NOT normally have a high tf-idf in a SOTU address -- to appear, with a source code barnacle attached.

'Crude' and 'freshman' are sort of interesting. I suspected 'crude' to be part of the phrase 'crude oil', (and indeed the 1975 SOTU has an extensive passage about crude oil). But the appearance of 'freshman' is likely an instance of a pet word or phrase (a word or phrase that a certain speaker tends to use a lot more than the average). As it turns out, President Ford had a particular affinity for the word, using it 3 times in 1975 and once in 1977. (Specifically, he liked the phrase "freshman congressman"). Since President Nixon also used that phrase (at least

once in a SOTU), and President Ford was originally his vice president, Ford likely picked it up from Nixon.

Affixes: In several speeches, the text parsing appears to have separated contractions into their component pronouns and verbal affixes. Such affixes appeared in the top results for multiple decades (e.g. 're' of "we're" or "you're", 've' of "I've" or "we've" or "you've", and 'm' of "I'm").

al qaeda: It is completely unsurprising that both 'al' and 'qaeda' are in the top results for the 2000s, especially considering the special post-9/11 SOTU. However, it is interesting to note that 'al' appears higher on the list. This could be because 'al' is simply the definite article in Arabic, and would therefore also appear in front of other Arabic words and names.

Party differences: The differences between the top results by party are not particularly informative. The encoding problems from the 70s (during which the president was republican for 8/10 years) have taken up a lot of space in the republican top 20. Words associated with terrorism and September 11th ('homeland', 'al', 'qaeda', 'regime', '2001', '11th') have also taken a starring role, though we cannot really know whether this is a result of republican values at work, or simply due to the fact that the only president represented during and after September 11th is George W. Bush. That is to say: if the president during that time had been democratic, would this event have been discussed so frequently as to become disproportionately represented in data coming from several different decades? We do not know.

Meanwhile, the results from the democratic speeches might offer more information in the way of 'party values', with words like 'diversity', 'scholarships', and 'empowerment'.

ought: The top democratic result, 'ought', is odd. As a modal auxiliary (a verbal helper that points to another possible version of the world), 'ought' should have been filtered out by the stoplist. However, it is a sort of archaic word (we usually use 'should' nowadays), and must not have appeared in the stoplist for that reason. Its archaism would also explain its evidently relatively low frequency in the corpus as a whole. All of this suggests that 'ought' is likely another pet word (like 'freshman'). As it turns out, again, this appears to be the case. Clinton evidently loved 'ought.' Of the word's 41 appearances in the corpus, 40 of them were by Clinton. This is the exact sort of linguistic idiosyncrasy that the stoplist was supposed to prevent against, but now I know for next time to include an even more comprehensive stoplist.

Final Thoughts

As a final task, let's look at the results as a whole. One of the most notable trends I see across decades is a tendency for high-ranked "conflict/enemy of the moment"-related terms. At least one such term appears in the top 10 of every decade except the 1970s (likely due to the encoding error). For the 1940s, the conflict is WWII: 'reconversion' appears because Truman liked the phrase "reconversion to a peacetime economy" after the war ended. For the 1950s, both 'communist' AND 'communists' appear, due to McCarthyism. For the 1960s, 'vietnam' is

literally the No. 1 result. For the 1980s, 'afghanistan' and 'nicaragua' appear within the top 5 due to US involvement in the Soviet-Afghan war and the Nicaraguan revolution. For the 1990s, 'bosnia' clocks in at No. 5 because of US involvement in the Bosnian war. And in the 2000s, of course, September 11th takes it all.

But why is this? Is it simply because the US is and always has been a militaristic nation? So American values dictate that the president must discuss at length whatever armed conflict we are currently in, or have just gotten out of, or have inserted ourselves into?

That is possible, I suppose, but I'd like to propose a more optimistic explanation: that the SOTU always mentions both good things and bad, but the good things are often the same across decades, so they don't show up as obviously in this type of analysis.