Final Project

Lautaro Cella

6/1/2022

Brief substantive background / goal

The Chilean party system has been highly stable since the country's democratization in 1990 (Valenzuela et al. 2018). It has been ordered around two coalitions that were formed to support or oppose dictator Augusto Pinochet's continued rule in the 1988 referendum. Both coalitions had ideological and programmatic tendencies: the Concertacion on the center-left, and the Alianza on the center-right. However, since massive protests erupted in the country in 2019 and a referendum was called to reform the Pinochet-sanctioned Constitution, both coalitions have lost most of its support. In 2020, independents with no partisan affiliations won the most seats in the Constitutional Assembly. In the 2021 presidential election, for the first time since democratization, the two candidates who received the most votes did not belong to the center-left or center-right coalition.

The literature predicts partisan identification to weaken when ideological polarization is low and coalitions become indistinguishable from one another (Lupu 2016). In this project, I will gather descriptive evidence to preliminary evaluate whether this hypothesis applies to the Chilean case. My expectation is that it does not. Center-right and center-left speeches will be different in ideological terms. I do not causally test the hypothesis since I only focus on ideology and do not look at partisanship.

I will analyze State of the Union speeches from Chilean Presidents between 1990 and 2021. Are speeches from center-left and center-right Presidents different in ideological terms? What words are associated with each coalition?

I will use Wordfish, an unsupervised learning method, to discover words that distinguish locations on a political spectrum. A party's position in the spectrum is assumed to affect the rate at which words are used in texts. I assume that State of the Union speeches capture positions on an ideological dimension. This is a justifiable assumption given the nature of the Chilean party system and the use of Wordfish to study State of the Union speeches in other countries.

Note: the Rmd. version of this file includes all my code, which I did not include in the PDF version.

Collecting data

First, I web scraped the website of the Chilean Congress to gather every State of the Union speech from 1990 to 2021. I used selector gadget to identify the date, the speaker, and the link with the speech. The speeches were in PDF format, so I used the function pdf_text to read them.

PDF text extraction was challenging. Luckily, most texts looked fine, they just needed to be cleaned. I realized that they had a lot of "/n/n" and "/n/nexampleword". Besides, the function failed to read the PDFs from 1996 and 1997. The quality of these two PDFs was very poor.

Cleaning / pre-processing data

The pdf_text function gave me a list of lists as an output, so I had to transform this element. I constructed a data frame with the text and document level variables (year, president, coalition). Using the data frame, I created a corpus with a meaningful identifier for each document.

One challenge was that the texts contained lots of "/n/nexampleword". I had to get rid of this using regular expressions. If I tried to delete this with the token function, I would end up with the letter "n" before many words "nexampleword" or "npresident".

I dropped documents from 1996 and 1997 because they were empty. I could not find the documents elsewhere. I also discovered that the document from 2010 was wrong, as the link contained a wrong PDF. It was not a State of the Union speech. Thus, I corrected this mistake and uploaded the right document, which was available online elsewhere.

I decided to keep the text in Spanish. All the pre-processing steps worked in Spanish. However, I noticed that stemming works worse in Spanish than in English.

I tokenized the text to unigrams. I removed punctuation, numbers, and symbols. I converted all characters to lowercase. I removed stop words in Spanish. I "stemmed" words. Then, I created a document term matrix.

I got rid of words that don't appear in 20% of documents. This is an important decision because I employ Wordfish and I only have speeches from one coalition at a given year. If the relevant political issues change across time and new vocabulary appears in year t+1, then this vocabulary will differentiate texts at year t and year t+1. So, I may pick up a policy agenda shift in texts, when I am interested in coalition ideological positions. By only keeping words that are mentioned in 20% of speeches, I am keeping words relevant enough to be mentioned over time by either one or both coalitions.

I plotted a word cloud with top words. After looking at it, I removed additional Spanish stop words. I also removed words that are too frequent and do not differentiate ideology among coalitions (Chile, law, government, president, etc.).

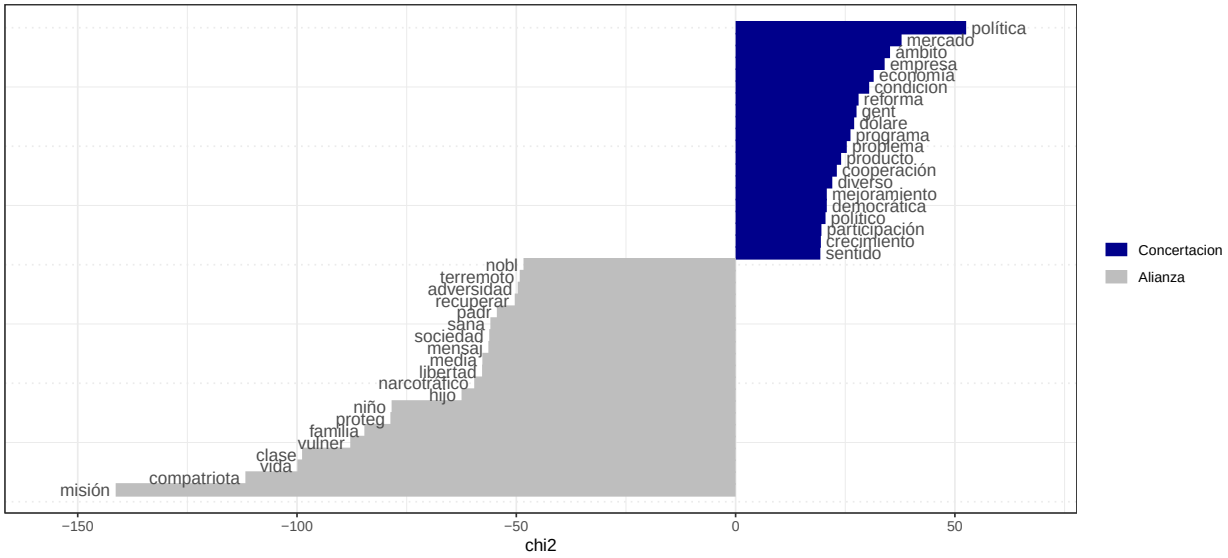
After all pre-processing steps, I was left with 30 documents and 4037 words.

Analysis and visualization

Distinctive Words

Using textstat_keyness(), I compared frequencies of words between center-left and center-right speeches.

```
textplot_keyness(key)
```



The center-left uses more words related to liberal economic development (market, company, US dollars, cooperation), and possibly university reform. The center-right uses more words related to family values (father, son, family), and narcotraffic.

The center-left uses more words related to democracy, probably because it governed during first years of democracy. The center-right uses "earthquake" more since it was in power after the 2010 earthquake.

Ideological Scaling: Wordfish

Wordfish is an unsupervised one-dimensional text scaling method that estimates the positions of documents solely based on the observed word frequencies. Reference texts are not required. It assumes the speaker has a position in a low-dimensional political space, which leads to the rate at which words are used. Word usage is independent of other words, drawn from Poisson distribution.

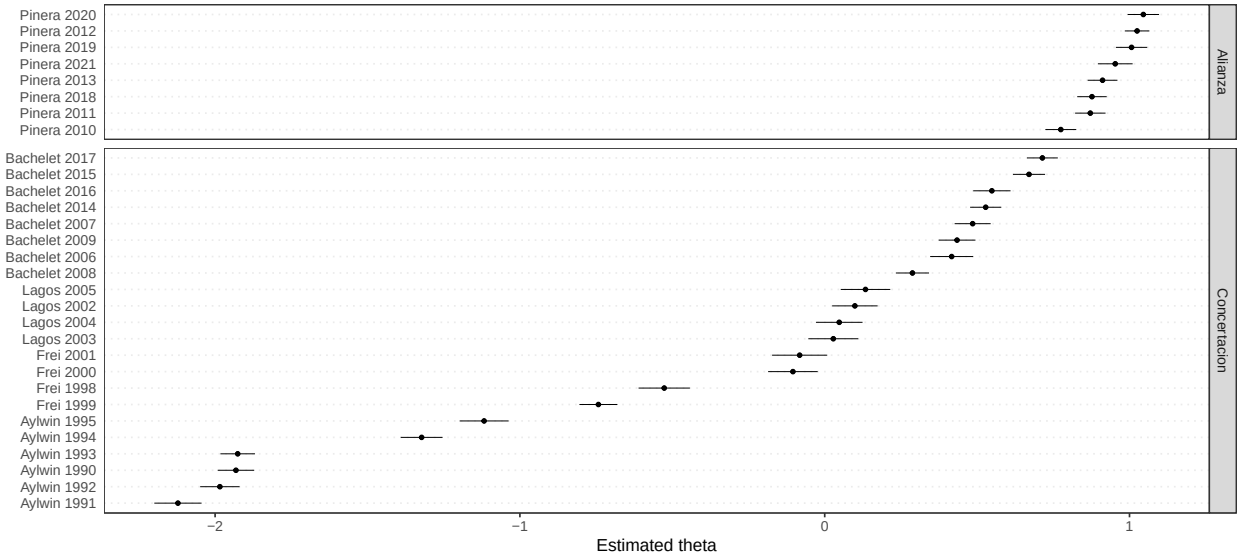
Thera are four parameters: word fixed effects (psi), word weigths (beta), document fixed effects (alpha), and document positions (theta).

```
# I force center-left Bachelet 2014 to have a smaller score than  center-right Pinera 2013
tmod_wf <- textmodel_wordfish(dfm_inaug_freq, dir = c(8, 9), tol = c(1e-06, 1e-08))
```

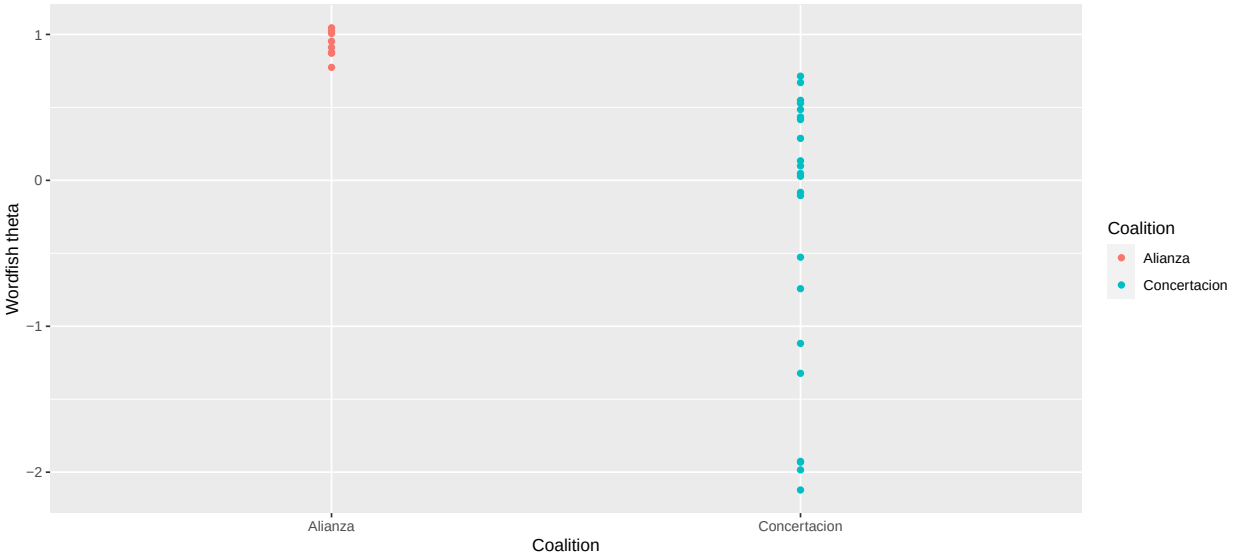
The theta scores estimate the ideological position of a text. I plotted the theta scores by political coalition of the speaker. The center-right texts are always to the right of the center-left texts.

I notice there is a time trend, with speeches moving to the right with time. I believe this may be because of variation in policy issues discussed, even though I got rid of words that don't appear in 20% of documents. The literature does not suggest that President Aylwin was more of a leftist than President Bachelet. In spite of this, I believe Wordfish captures an ideological difference between speeches and not just different word usage across time.

```
textplot_scale1d(tmod_wf, groups = dfm_inaug_freq$coalition)
```

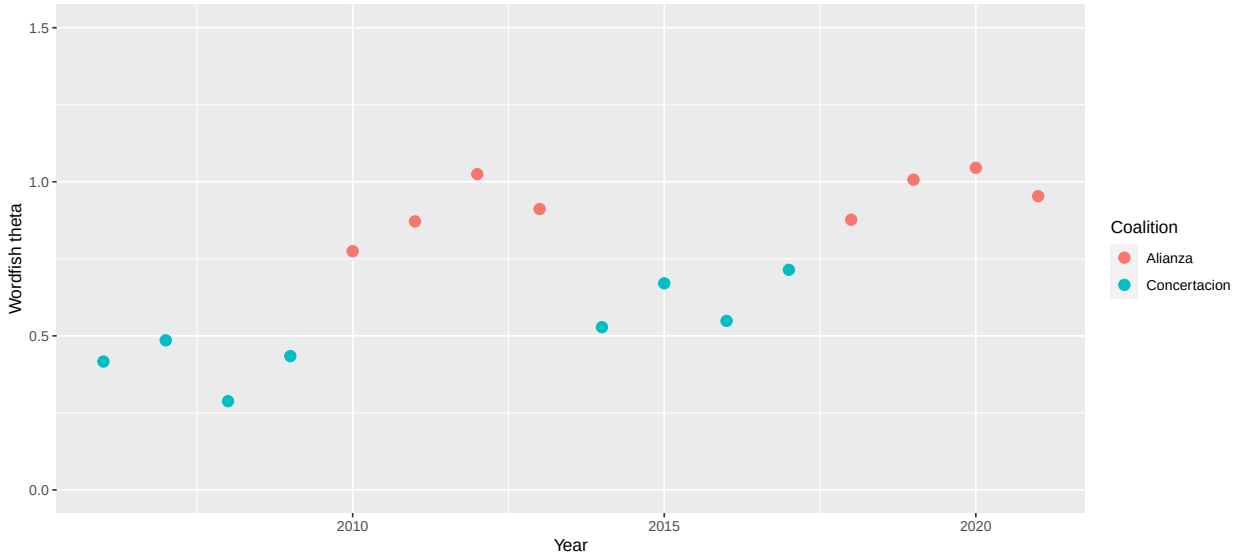



```
plot2
```



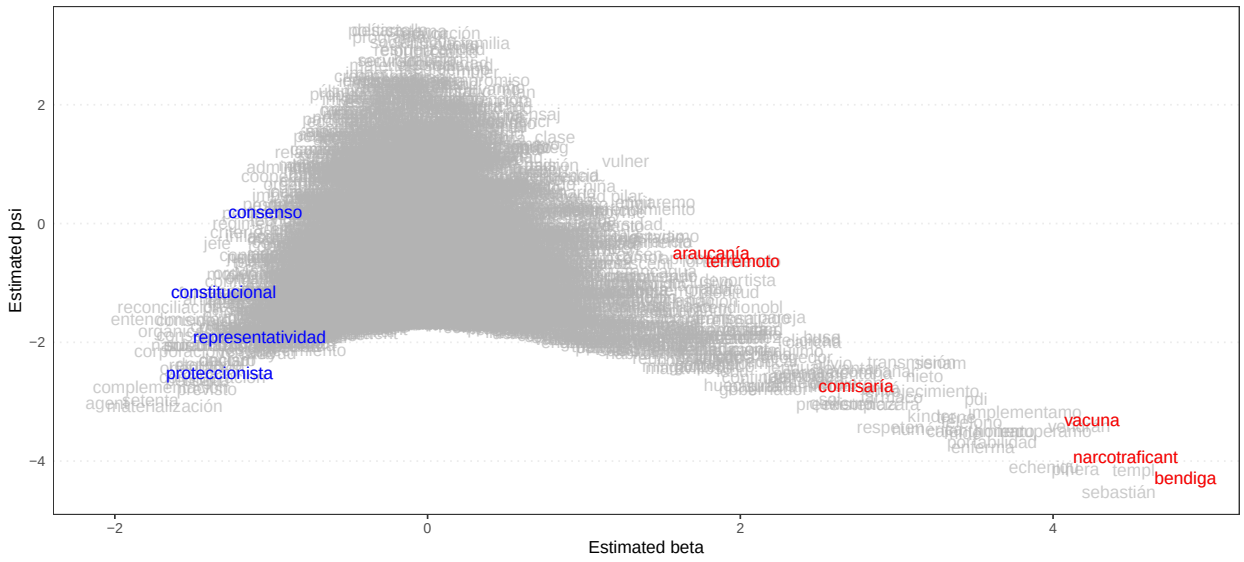
I also plotted the scores between 2006 and 2021, to show that scores are not completely determined by the time trend. This period includes two mandates by center-left Michelle Bachelet and two mandates by center-right Sebastian Pinera. I am holding issues more or less constant by restricting the time period. The plot shows that there are substantive differences in ideology, identified by words usage. The center-left President always has a lower score than the center-right one.

```
plot3
```



Finally, I plot the estimated word positions and highlight certain features. The beta coefficient measures the importance of a word in discriminating the underlying position. Words with a high absolute value discriminate well. The psi coefficient is a word-fixed effects. Words with a higher value are more common.

```
plot4
```



The center-right uses words related to security (narcotraffic, police station) and the indigenous conflict in the South (Araucania). It also uses the word "bless", usually in the context of "God bless Chile". The more secular center-left does not use this phrase. The center-left uses words related to the Constitution (potentially referring to a constitutional reform) and economic proteccionism.

Finally, the center-right uses words related to the pandemic (vaccine) and earthquake. This is probably because it governed during the pandemic and post the 2010 earthquake. The center-left uses words related to democratization (consensus, reconciliation). This could be because it governed during that time.

Future work

In this exploratory and descriptive work, I have analyzed the ideological dimension of State of the Union speeches in Chile. My study suggest speeches from the two main coalitions are different ideologically. Yet, I have not looked at the relationship between ideology and partisanship. I have also not answered whether speeches have become more or less similar in ideology across time. To conduct this other analysis, I would need speeches from the different coalitions from the same year.

This Summer, I will go to Chile for fieldwork. Among other things, I will gather campaign speeches from presidential candidates from the two coalitions in every presidential election between 1990 and 2021. Once I have collected these speeches, I will replicate the analysis using the new corpus. This way, I would be able to assess whether coalitions moved together over time.

Alternative texts that I could use are party manifestos and legislative speeches. The problem with party manifestos is that the literature argues that, unlike in Europe, in Latin America they are not as useful to analyze ideology. The challenge I encountered with legislative speeches is that online documents containing them change structure across documents and across time. Legislative projects and speakers are not identifiable with a uniform structure.

If I find out a way to work with legislative speeches, I will conduct a study where each intervention made by a Senator in a given Senate session is the unit of analysis. My measurement strategy would be to use the accuracy of machine classifiers (the proportion of correct predictions of the Senator's coalition in a year/legislative period). The labels would be the coalitions of Senators, predicted from their speeches. When the learner discriminates members well, the period would be one of high polarization. When accuracy of the classifier is low, polarization would be low. I would use different classifiers in my analysis.