Register analysis of Chilean presidents speeches over 30 years*

Pamela Inostroza Fernández

Master in Statistics 2020/2021

Abstract

Factor analysis in addition to Correspondence Analysis was performed to identify characteristics of the last 31 speeches of Chilean presidents. The attributes of each speech, such as the number of words of different types, numbers, and the use of specific adverbs were used to identify how the structure of presidential speeches has changed accordingly with important events, like the change of the century, the first female president, the first right-wing president, and some catastrophes like 2010's earthquake and 2020's pandemic.

*All codes and txt files available github.com/PamelaInostroza/MCL. **Current version**: January 10, 2021; **Corresponding author**: inostroza.f.pamela@gmail.com.

1 INTRODUCTION

Every year, presidents had to communicate to the National Congress their main achievements and their legislative priorities for the near future. There is multiple information available for this type of texts. For this research, Chilean presidency transcripts will be used. The main objective will be to visualize some insights into the evolution of the linguistic used by each president according to the main events the country faced at the time.

Chilean speeches were chosen because they have a recent history that can help to identify some patterns in linguistic use. Data from the year 1990 when Chile got back to democracy from a dictatorship that took place for 17 years.

In the latest history of Chilean presidents, there have been 5 different presidents, 2 of them have repeated their mandate. In total 7 governments. This research will be focused in identify if the way of referencing to the congress/people in their speeches show specific characteristics.

Since the presidential function definitely has changed since the back of democracy it is expected that the speeches show this change and evolution of words and the way to address the main problems that the country faces at the time.

In this study it will be assumed that the president is the author of their speech (even though this is not literally), this research will not be focused in the team behind the writing of the speeches, this can lead to more in-depth research.

In order to identify the degree in which every group of speeches are different from one another, correspondence analysis will be performed. Here, many features will be obtained from each speech, some of these features are related to the length of the speech and frequencies of specific types of words. Also, it will include characteristics such as frequency of positive, negative words, bigrams, trigrams and other characteristics.

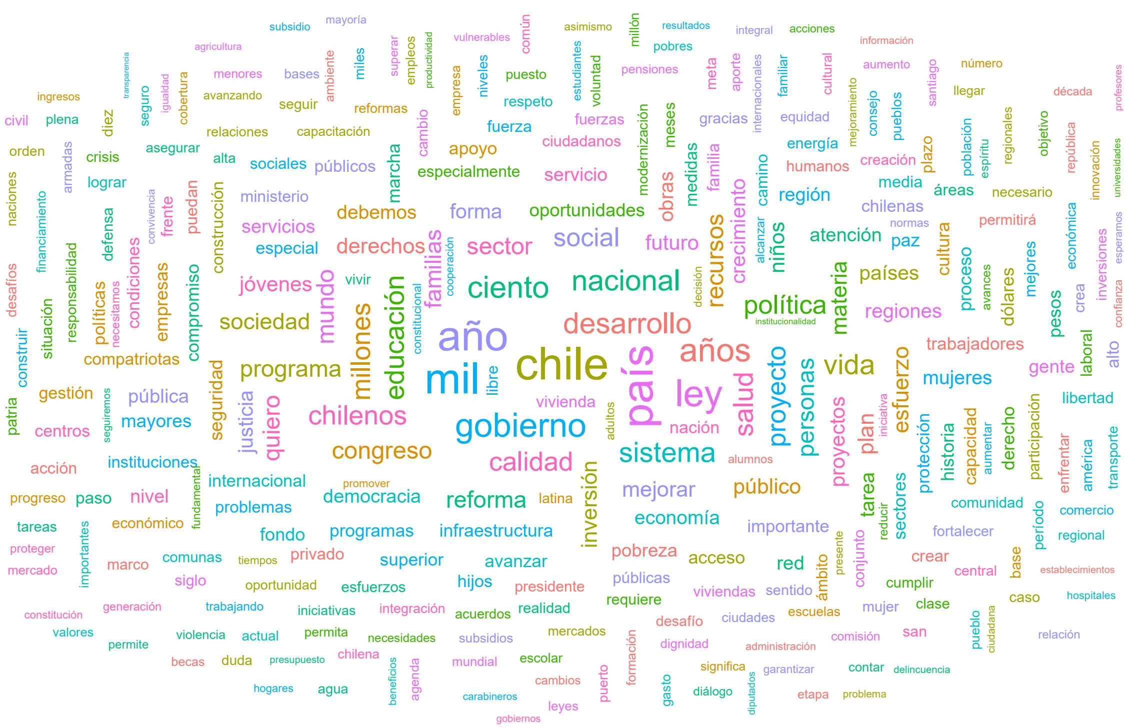
This research is performed completely on R software where libraries such as *dplyr* for the management of the data, *mclm* for specific functions for analysis of corpus linguistics, *ca* to perform correspondence analysis, *factanal* for exploratory factor analysis, *knitr* to create a reproducible report, *kableExtra* for tables visualization and *ggplot*, *ggwordcloud* and *ggpubr* for plots visualizations.

2 DESCRIPTION OF DATASET

The governmental speeches selected were produced in the same context from 1990 to 2020, 31 speeches in total. May 21th was established as the date when this public account speech should be performed every year. From the year 2017, this date was changed to June 1st[1].

Each speech is available in plain text (.txt), this files can be downloaded from the site indicated previously, these files were read using *get_fnames* function from **mclm** package in R. For the use of Spanish language it is important to specify the encoding as "UTF-8", as the corpus contains the special character "ñ" and tildes in most vowels (such as á,é,í,ó,ú). From these files is this possible to extract all words used.

As a first approach to the data, a global cloud of words is performed, Figure **??** shows the most used words in all speeches studied, the size of the word in the graph indicates how many times the word was found in all sets of speeches. Words at the middle of the plot indicate the most used words, most visually identifiable words are "Chile", "year", "country", "government", "law".



---

[1]This date was changed exceptionally for the year 2020 regarding the global pandemic to July 31st

Table 1: Name and period of each president

N	Periodo	Name
1	1990-1993	Patricio Aylwin
2	1994-1999	Eduardo Frei
3	2000-2005	Ricardo Lagos
4	2006-2009	Michelle Bachelet
5	2010-2013	Sebastian Pinera
6	2014-2017	Michelle Bachelet
7	2018-2020	Sebastian Pinera

The names read are saved for each year and president/mandate as a grouping variable identified by the option *requested_group = 1* from *re_retrieve_first()* function, the table 1 indicate all the periods studied and the president in charge on those years.

It will be interesting to see how different attributes change over time, as an example, we can see in figure 1 the number of tokens and types found in each set of speeches. The longest speech was in 1997 and the shortest in 2001. The third period of Lagos as president shows the short amount of tokens compared to the rest of the speeches. In the plot also can be seen the most frequent type, for the first years of democracy, *government* (gobierno) was most often used, then *country* (país), the longest speech has *development* (desarrollo) as the most frequent type. All following year use *Chile* as principal type word, last 7 years *country* is again the most common.

Much other information can be extracted from the corpus, as shown in figure 2, the proportion of use of some attributes such as doubt adverbs, numbers, types and verbs is very different each year, and in some cases clear decrease/increase in their use.

This information is very interesting but not very helpful by itself in order to identify common patterns from every speech. For this reason, further analysis is needed, that allows us to combine all this information and obtain consolidate results about patterns in the speeches.

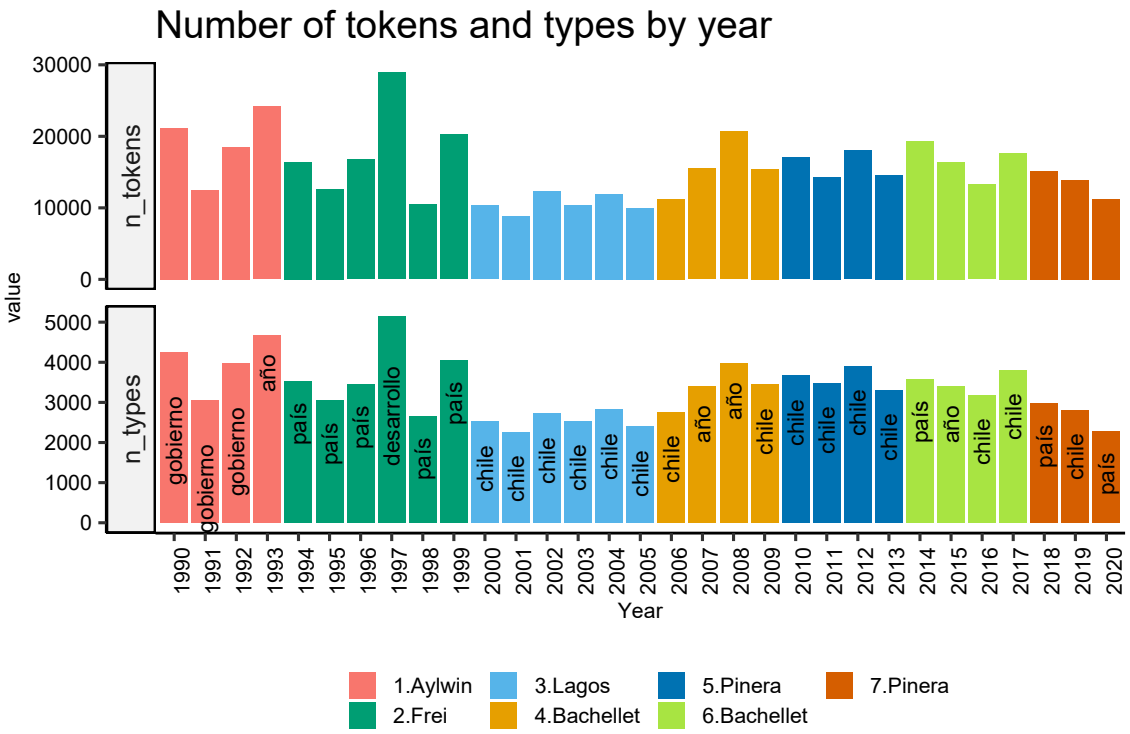


Figure 1: Number of tokens and types in speeches

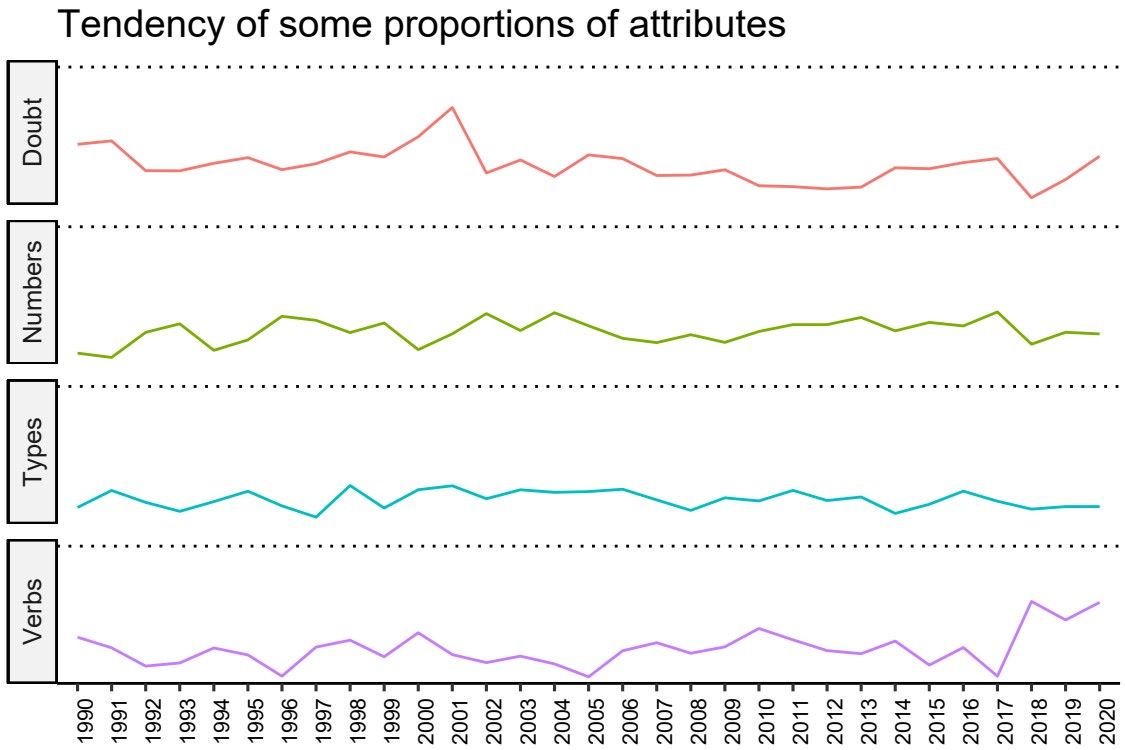


Figure 2: Tendency (proportion of use) of speeches attributes

3 ANALYSIS

The aim of this study is to review in which degree president speeches are related in the sense of similarities or dissimilarities. For this, many features will be extracted from the different corpus in order to obtain characterization that helps to perform this comparative analysis.

The attributes used to the reduction of dimension is listed as follows:

- word_len: average word length (expressed as the number of characters)
- p_dig: the proportion of numbers used
- p_types: the proportion of types by total token [2]
- p_bigr: the proportion of bigrams used [3]
- p_trigr: the proportion of trigrams used [4]
- p_stopw: the proportion of stop words used
- p_funct: proportion of functional words
- p_verb: the proportion of verbs used
- p_mode: the proportion of modal adverbs *(i.e., good, bad, simply, sincerely, fast)*
- p_advtim: the proportion of time adverbs *(i.e., now, before, after, later, soon, yesterday, earlier, right now, yet, today)*
- p_advlug: the proportion of place adverbs *(i.e., here, there, near, far, out, inside, around, aside, in, behind, front)*
- p_advqua: the proportion of quantity adverbs *(i.e., much, a lot, very, almost, all, nothing, some, medium, more, less, also)*
- p_neg: the proportion of negative words *(i.e., no, neither, never, ever, nobody, nothing, no one)*
- p_pos: the proportion of positive words *(i.e., yes, too, some, always)*
- p_dou: the proportion of doubt adverbs *(i.e., might, maybe, perhaps)*

3.1 Correspondence analysis

Correspondence analysis is a dimension reduction technique that allows us to create relations among rows, columns and between rows and columns. This type of analysis is very flexible in the sense that no extra restrictions are imposed on contingency tables.

The main idea resides in a dimensional reduction, where each observation is represented by a point and each attribute as a dimension. This representation can be visualized in a two-dimensional space where a cloud of points indicates the relation between new dimensions.

[2] A graphical representation of most used types can be found in the annex, figure **??**

[3] A graphical representation of most used bigrams can be found in the annex, figure **??**

[4] A graphical representation of most used trigrams can be found in the annex, figure **??**

Table 2: Variance explained by each dimension

	Eigenvalue	Variance explained	Cumulative Variance explained
Dim.1	0.022	49.2	49.2
Dim.2	0.011	23.4	72.6
Dim.3	0.007	15.7	88.3
Dim.4	0.003	7.3	95.6

As indicated in table 2, 72.6% of the total variance can be explained by reducing all the attributes in 2 dimensions, figure 3 shows how these 2 dimensions are distributed. It can be seen that most of the attributes are together in the II quadrant (upper left), from here we can say that is possible to reduce the dimensions even more in order to facilitate the interpretation. For this reason, factor analysis is also performed in order to reduce the number of attributes (15) into a few factors (4) that summarize all similar attributes.

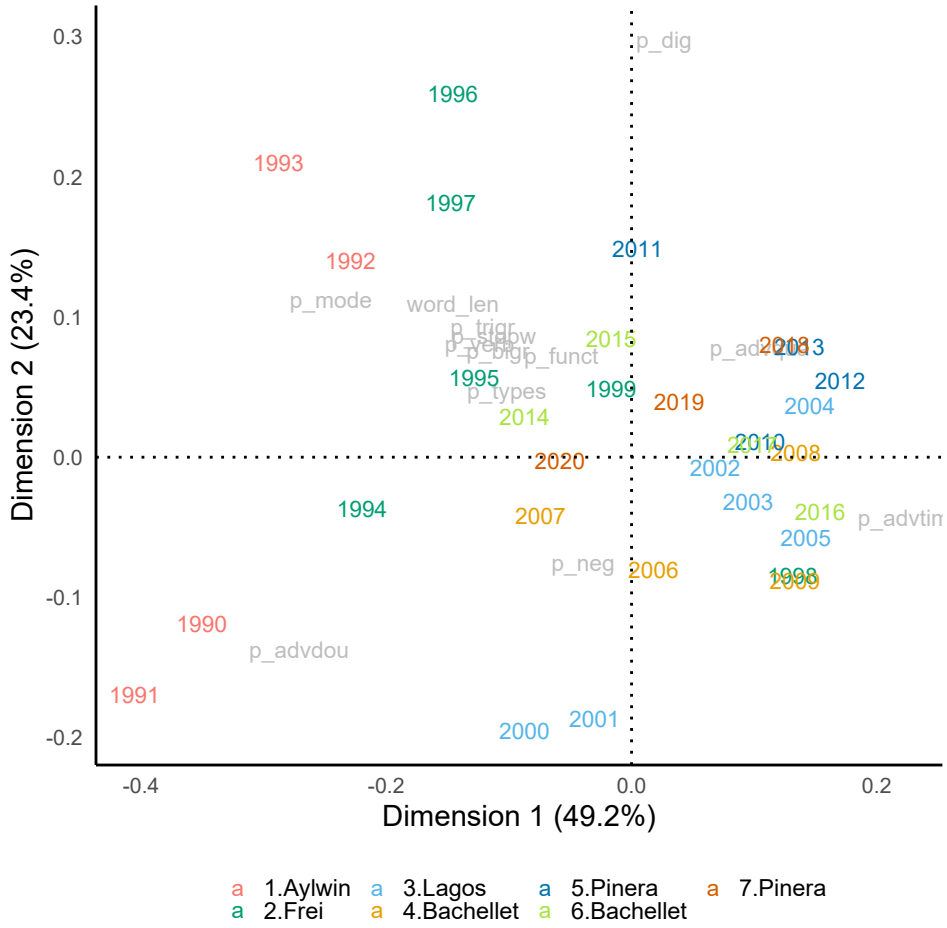


Figure 3: Correspondence analysis of attributes

3.2 Factor analysis

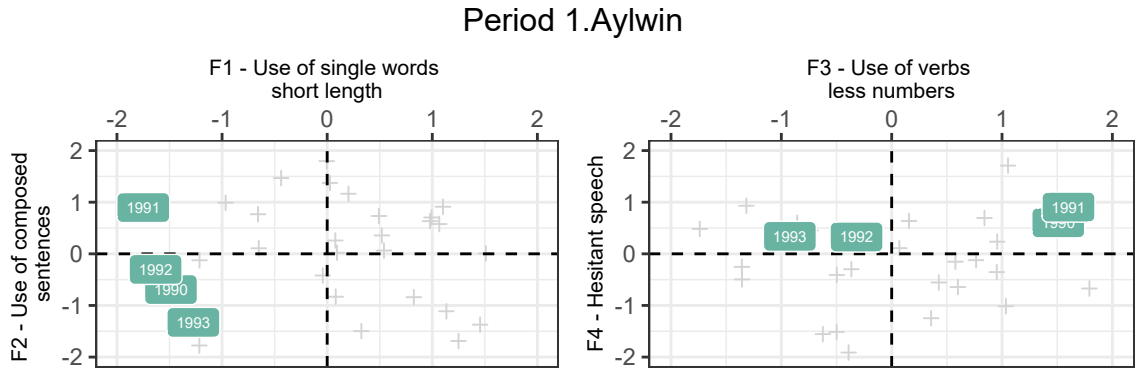
After performing factor analysis, it is possible to indicate that attributes such as the proportion of stop and functional words, time, quantity and modal adverbs, negative words are strongly related, hence, it is possible to combine them and obtain a new factor. It is important to mention that when the loading is negative, as for modal adverbs and length of words, this means that the attribute should be read inversely. This factor could be named as **Single words, short length**. As Factor 2, can be combined attributes related to the proportion of types, bigrams and trigrams, named as **Composed sentences**. Factor 3 is mainly composed by the proportion of verbs and digits, it was called, **Use of verbs and no numbers**. Lastly, factor 4 include doubt adverbs, it was called **Use of hesitant words**. With this reduction of attributes, that explain more than 80% of the total variance, it is possible to redo the previous graph where the relations between the attributes and each year can be clearly differentiated by this four interpretable factors.

Table 3: Factors identified

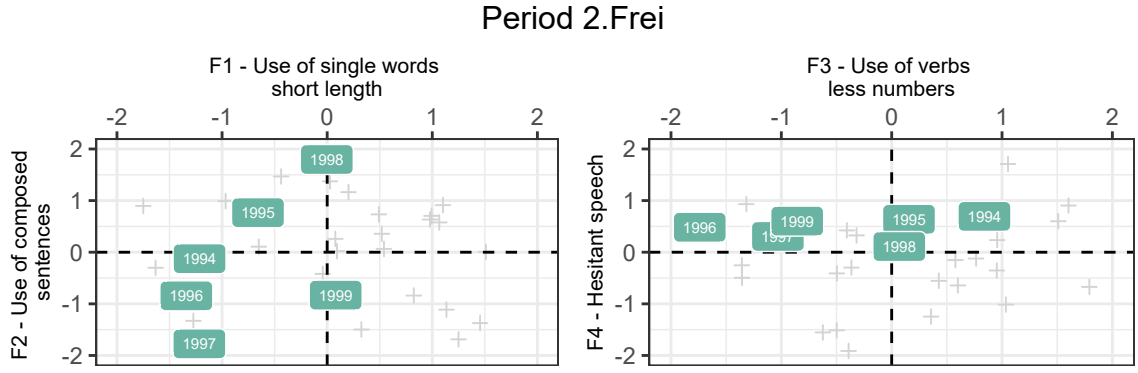
Attribute	F1-Use of single words, short lenght	F2-Use of composed sentences	F4-Use of verbs but less numbers	F5-Use of hesitant adverbs
p_funct	0.880			
p_stopw	0.867			
p_advtim	0.690			
p_advqua	0.658			-0.589
p_neg	0.536			
p_mode	-0.618			
word_len	-0.960			
p_bigr		0.980		
p_types		0.946		
p_trigr		0.941		
p_verb			0.702	
p_dig			-0.875	
p_advdou				0.756

With this graphical visualisation it is possible to identify that first two periods and social democrats presidents, figures 4a and 4b *(1.Aylwin, 2.Frei)*, in F1-F2 biplot are concentrated in the left side of the plot, which corresponds to longer speeches and use of single words. Similar behaviour for both periods is found in F3-F4 biplot, in both cases, their speeches are concentrated in the upper side, which means more use of doubt adverbs.

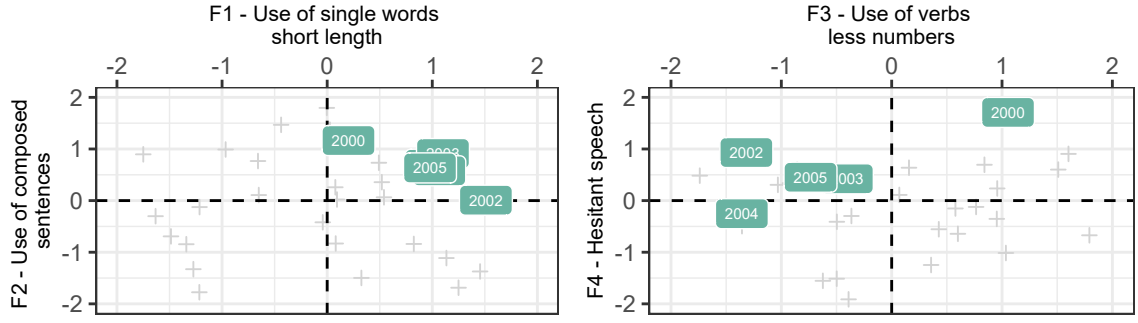
During the third period, figure 4c *(3.Lagos)*, the first socialist president after return to the democracy, another distribution can be seen in the graphs compared to the previous ones, as in the first biplot F1-F2, speeches are concentrated in the upper right quadrant which means shorter than other speeches but use more single words, types, bigrams and trigrams. In the second plot, speeches are mainly in the upper left quadrant, this means that the use of doubt adverbs was common as previous presidents but after the year 2000 the use of verbs decreased and the use of numbers increased.



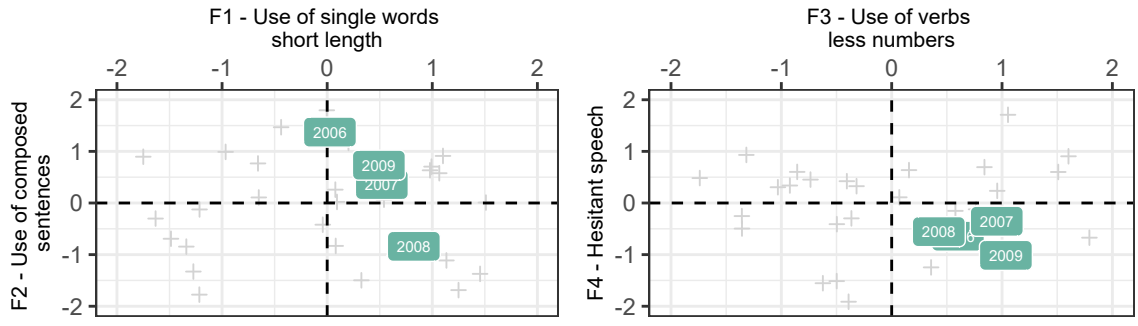
(a) Biplot 1.Aylwin



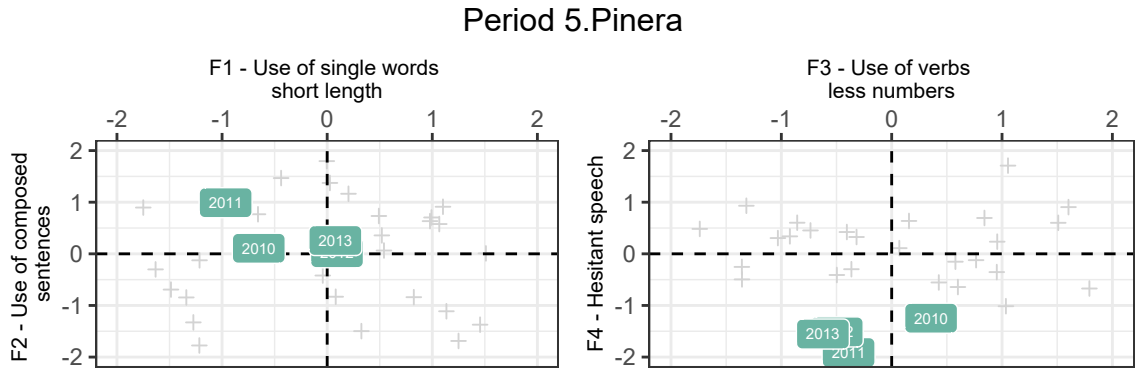
(b) Biplot 2.Frei



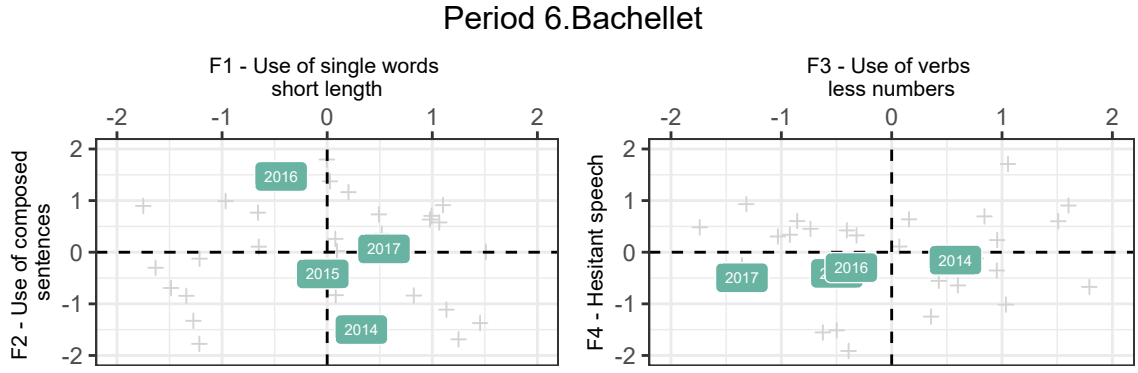
(c) Biplot 3.Lagos



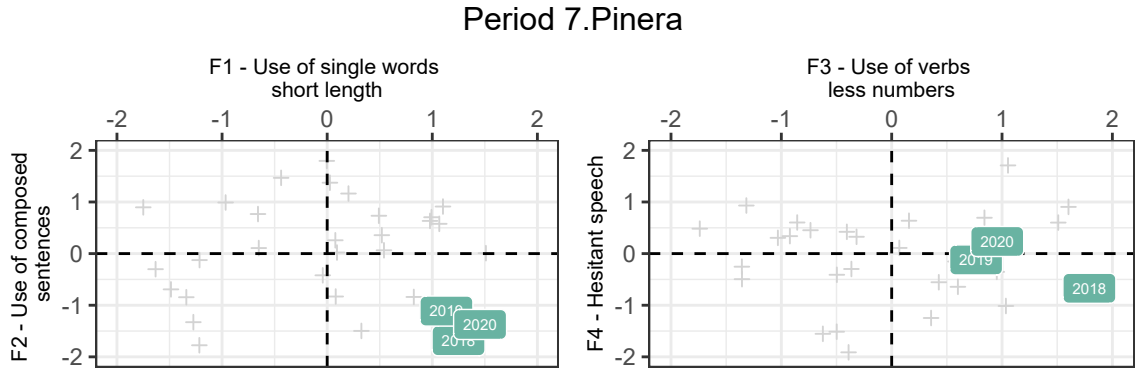
(d) Biplot 4.Bachellet



(e) Biplot 5.Pinera



(f) Biplot 6.Bachellet



(g) Biplot 7.Pinera

Figure 4: Correspondence analysis of 4 main factors

Continuing the tendency from the previous period, in figure 4d, fourth period endorsed by the second socialist and first female president *(4.Bachellet)*, her speeches are concentrated in the upper right quadrant, maintaining shorter speeches with use of single words and more types and bigrams/trigrams with exception of the year 2008, where fewer types/bigrams/trigrams where used. For the second F3-F4 biplot, the tendency changed dramatically as these speeches are mainly concentrated in the first quadrant, which means that fewer doubt adverbs were

used, and more verbs and fewer digits.

In the fifth period *(5.Pinera)*, the first right-wing president since the back to the democracy, from national renovation party. In figure 4e the distribution of his speeches are almost right in the middle in F1-F2 biplot, this means that the length of the speeches is medium compared to the other speeches, and also in the use of single terms and composed types, bigrams and trigrams. In the second biplot (F3-F4), can be seen that no doubt adverbs are used compared to the other speeches, fewer verbs and more numbers are used after the first year of his mandate 2010 when the country suffered a massive earthquake.

For the second period of *(6.Bachellet)*, figure 4f shows that even though the position in the middle of the plot is maintained as in her first period (F1-F2 biplot), this time is more disperse (in 2014 less use of composed words and in year 2016, more use of composed words). For the second biplot F3-F4, even though the use of doubt adverbs is lower, after the first year, more numbers are used and fewer verbs compared to other speeches.

Finally, the last period studied *(7.Pinera)*, and the second period for the only right-wing president, figure 4g shows the shortest speeches of all in F1-F2 biplot, with more use of single words, and less use of types, bigrams/trigrams. In the second biplot F3-F4, the tendency of no use of doubt adverbs was broken on the year 2020 when the COVID-19 pandemic occurs and more verb is used and fewer numbers in the three years.

4 CONCLUSION

Visually, it is easy to identify the main differences between characteristics of the speeches of each president, and thanks to the reduction of dimension it is easier to interpret.

As a conclusion, without going into any deep political analysis, the attributes selected from each speech, allow us to identify the main behaviour of the way the speeches were performed based on the period they were given.

It is clear the caution that is used in the wording until the year 2000 (first 10 years of democracy), maybe this change was because of the century change and the new technological development starting to be more accessible.

Another important event can be seen when the first woman president was elected, even though the previous president was from the same party, no relation can be found in the characteristic of their speeches.

The most obvious change can be seen when the first right-wing president was in charge, sadly the very first year he started, a massive earthquake (8.8Mww) occur in Chile and this can be seen in the analysis too.

Finally, the last period, the lowest approval rating president show a completely different characterization of his speeches, even different from his previous mandate. Also, the cautious wording is evidenced for the year 2020 when the pandemic occurs.