



```
library(dplyr)
```

```
rladies_global %>%  
  filter(city == 'Santiago')
```

# Análisis de textos con R

Búsqueda de patrones lingüísticos  
en discursos políticos

Riva Quiroga



@rivaquioga

¿Qué nos interesa a los lingüistas?

**hacernos preguntas  
sobre los usos del  
lenguaje**

**En estos momentos estoy desarrollando un proyecto que busca identificar patrones en la forma en que se usa el lenguaje en los mensajes presidenciales (cuenta pública)**

Por ejemplo, ¿quién dijo esto?

**“En el nivel de la enseñanza superior, nuestro principal problema durante décadas fue el de la equidad”**

**(Suena a Bachelet 2017)**

(Pero no)

**“En el nivel de la enseñanza superior, nuestro principal problema durante décadas fue el de la equidad”**

**Frei  
1998**

**¿Hace 20 años que se habla de lo mismo?**

**Los mensajes  
presidenciales son un  
género discursivo**

Entre otras cosas, esto implica que quienes los enuncian tienen que abordar ciertos temas para que el discurso cumpla su propósito.

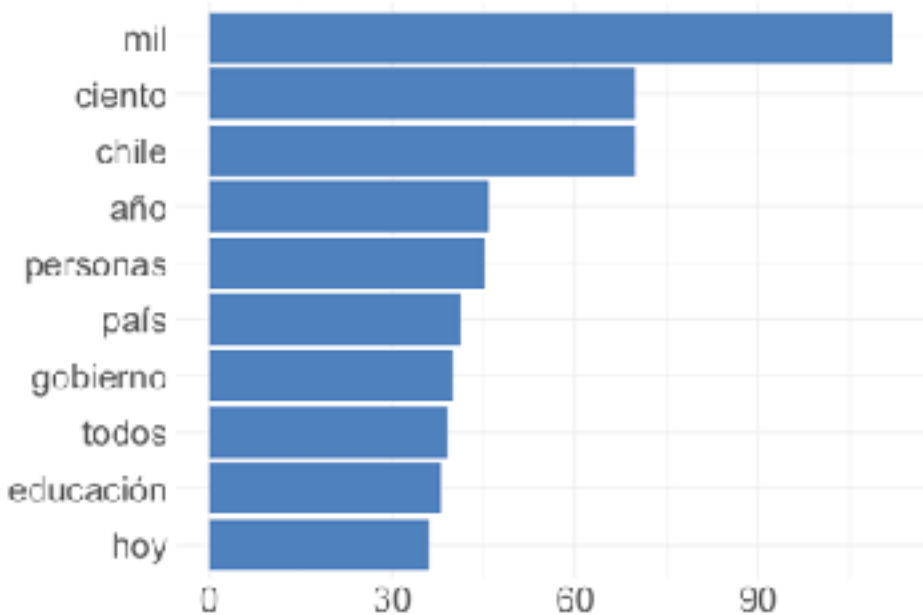
Sus opciones en ese ámbito son limitadas: no pueden hablar de lo que les dé la gana.

**Hagamos algo habitual:  
analicemos las palabras  
más frecuentes en algunos  
mensajes**



```
library(tidytext)
df_2017 %>%
  unnest_tokens(word, text) %>%
  anti_join(stopwords_es) %>%
  count(word, sort = TRUE)
```

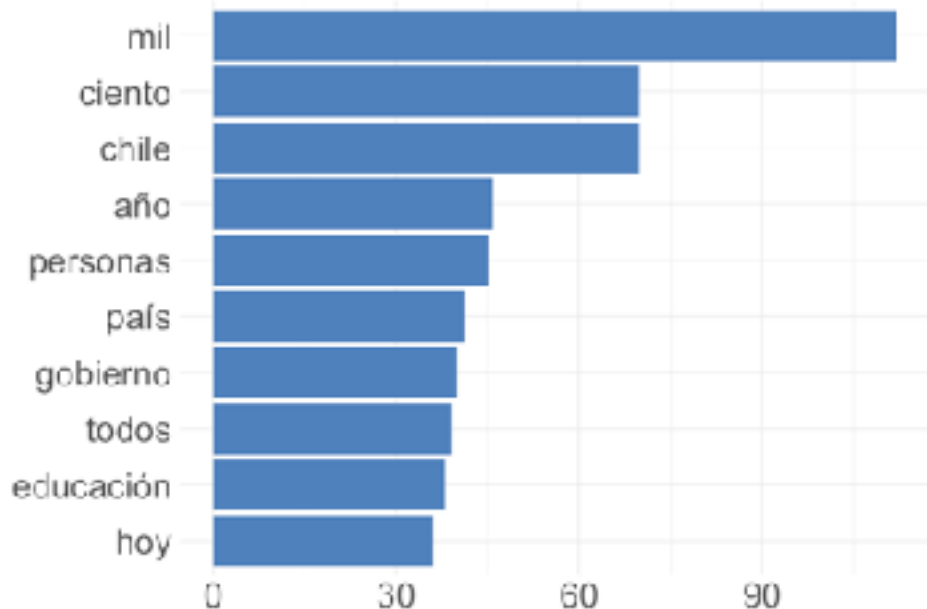
# 2017



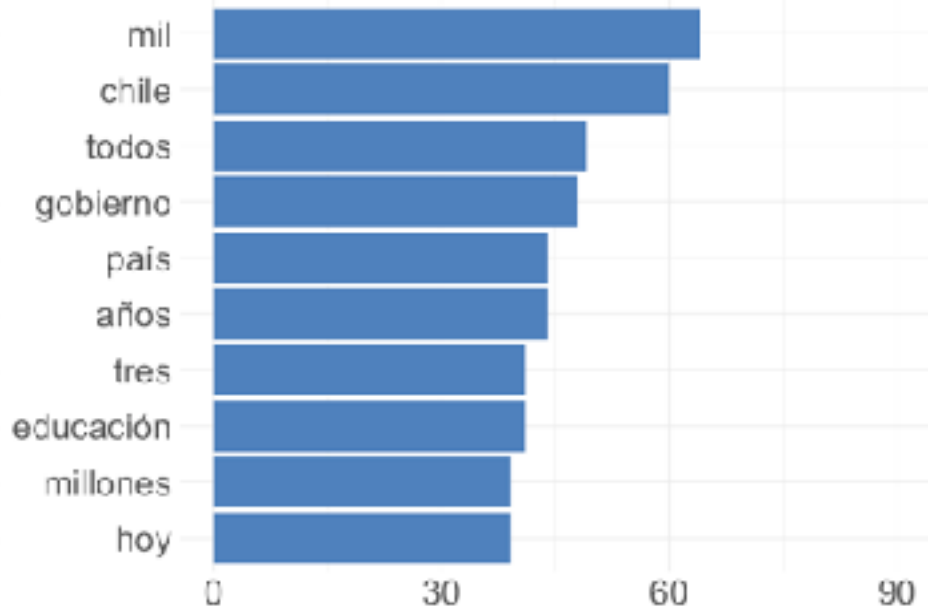
**Estos son los términos más usados en el último discurso de Bachelet. Todo dentro de lo esperado.**

**¿Qué pasa si analizamos un discurso de una orientación política distinta?**

# Bachelet 2017

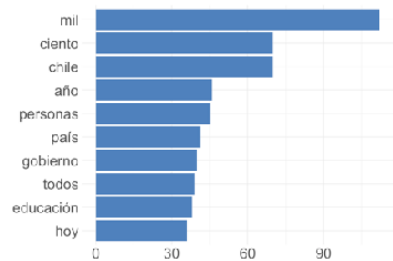


# Piñera 2013

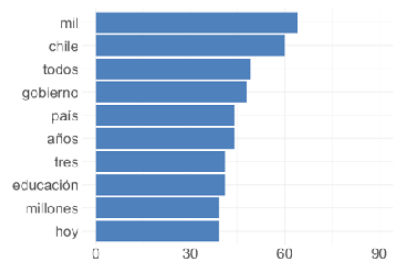


¡Son casi las mismas!

2017



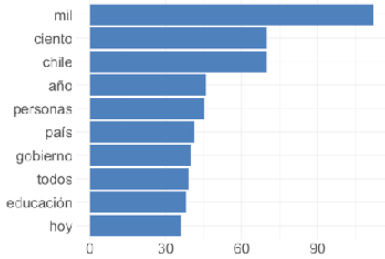
2013



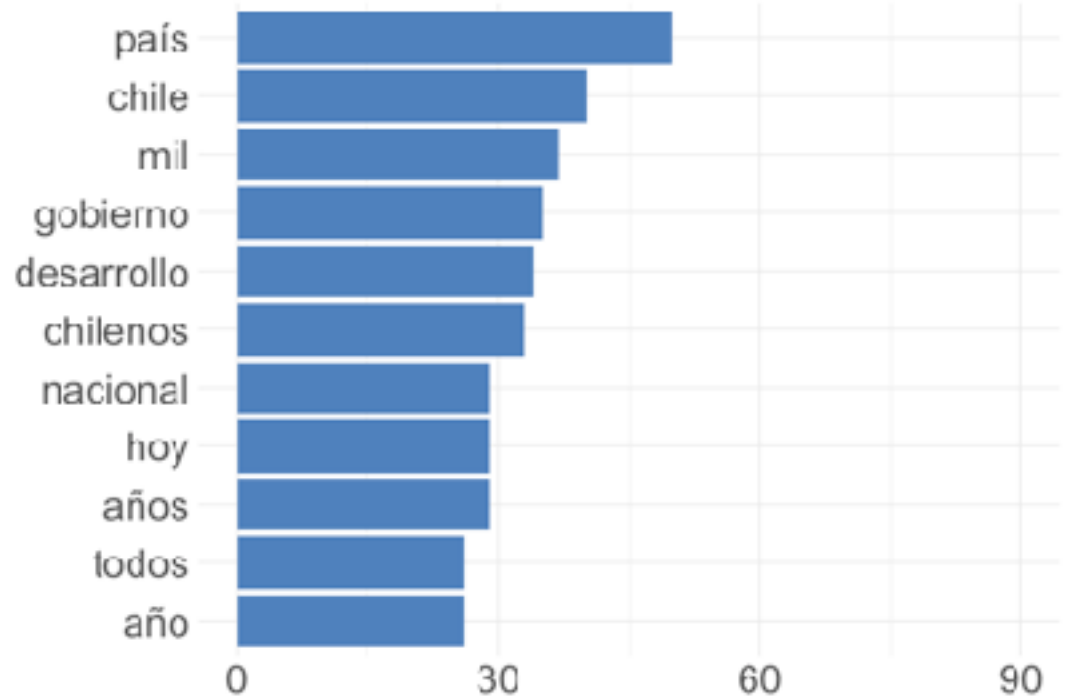
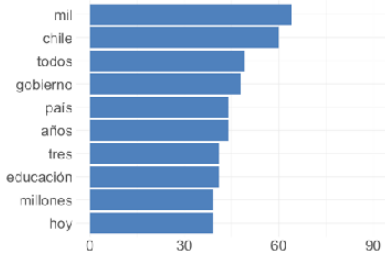
Hagamos una  
comparación más  
extrema

# Pinochet 1989

## Bachelet 2017



## Piñera 2013



¡Son casi las mismas!

# ¿Por qué pasa esto?

Todas esas palabras son términos casi indispensables en un mensaje presidencial. No hay opción: es necesario utilizarlas con frecuencia para que el mensaje cumpla su propósito.

**Antes de seguir, hagamos  
una distinción importante**



**hay dos enfoques principales para abordar  
el análisis de textos**

ignorar el orden y la función  
de las palabras



**“bolsa de  
palabras”**

```
bachelet_2015 %>%  
  unnest_tokens(word, text) %>%  
  anti_join(stopwords_es) %>%  
  count(word, sort = TRUE)  
  
# A tibble: 3,317 x 2  
  word      n  
  <chr> <int>  
1 año      98  
2 mil      70  
3 país     55  
4 chile    52  
5 ley      52  
6 todos    52  
7 proyecto 44  
# ... with 3,307 more rows
```

los textos se analizan a partir de tablas de frecuencia

**“bolsa de  
palabras”**

```
library(tm)
```

	Terms				
Docs	abarcas	abismo	bondad	buscar	...
1	0	0	0	0	
2	0	6	2	0	
3	0	1	0	0	
4	1	2	0	0	
5	0	1	0	1	
6	0	6	0	0	
...					

o matrices

“bolsa de  
palabras”

es lo que suele conocerse como  
**text mining**



**“bolsa de  
palabras”**



“bolsa de  
palabras”

el otro enfoque

**análisis y  
etiquetado  
sintáctico**

para el análisis sí importa el  
orden y función de las  
palabras



“bolsa de  
palabras”



análisis y  
etiquetado  
sintáctico

es lo que suele hacerse desde la  
**lingüística**



**“bolsa de  
palabras”**

**análisis y  
etiquetado  
sintáctico**



# “you shall know a word by the company it keeps”

(Firth 1957)



“bolsa de palabras”

análisis y  
etiquetado  
sintáctico

**es importante, entonces,  
no descartar el orden**

para eso podemos ocupar  
`library(cleanNLP)`

```
library(cleanNLP)
init_tokenizers(locale = "es")
tokens_2017 <- run_annotators(d2017v, as_strings = TRUE) %>%
  get_combine()
```

# A tibble: 19,632 x 6

	id	sid	tid	word	cid	spaces
	<int>	<int>	<int>	<chr>	<int>	<dbl>
1	1	1	1	Queridos	1	1
2	1	1	2	compatriotas	10	0
3	1	1	3	,	22	1
4	1	1	4	vengo	24	1
5	1	1	5	a	30	1
6	1	1	6	dar	32	1
7	1	1	7	cuenta	36	1
8	1	1	8	ante	43	1
9	1	1	9	mis	48	1
10	1	1	10	conciudadanos	52	1

# ... with 19,622 more rows

```
tokens_2017 %>%  
  filter(sid == 10)
```

```
# A tibble: 27 x 6
```

	id	sid	tid	word	cid	spaces
	<int>	<int>	<int>	<chr>	<int>	<dbl>
1	1	10	1	Un	1428	1
2	1	10	2	Chile	1431	1
3	1	10	3	con	1437	1
4	1	10	4	un	1441	1
5	1	10	5	malestar	1444	1
6	1	10	6	profundo	1453	1
7	1	10	7	por	1462	1
8	1	10	8	tantas	1466	1
9	1	10	9	barreras	1473	1
10	1	10	10	invisibles	1482	1

```
# ... with 17 more rows
```

**podemos filtrar  
por ubicación:  
por ejemplo, la  
oración 10.**

Este paquete también permite utilizar dos etiquetadores sintácticos para el español. Así podemos conocer la **función** que cumple cada palabra en el texto.

**spaCy**

(Python)

**coreNLP**

(Java)

**spaCy**  
(Python)

[Ocuparemos esta opción  
porque a veces instalar RJava  
se vuelve complicado, ¿no?]



```
reticulate::use_python("~por/aquí/llego/a/pyhton")  
reticulate::py_config()
```

**[No olvide esto al principio de su script,  
para que spaCy funcione  
correctamente]**

```
reticulate::use_python("~por/aquí/llego/a/pyhton")  
reticulate::py_config()
```

```
library(cleanNLP)  
init_spacy(model_name = "es")
```

```
nlp_2017 <- run_annotators(d2017v, as_strings = TRUE) %>%  
  get_combine()
```

**[Si el texto es muy largo esto puede demorar un poco. Sea paciente]**

```
nlp_2017 %>%  
  filter(sid == 10) %>%  
  select(word, lemma, upos, pos)
```

**Volvamos a mirar la oración 10, ahora con el análisis sintáctico.**

# Mucha información :)

```
nlp_2017 %>%  
  filter(sid == 10) %>%  
  select(word, lemma, upos, pos)
```

```
# A tibble: 27 x 4
```

	word	lemma	upos	pos
	<chr>	<chr>	<chr>	<chr>
1	Un	un	DET	DET__Definite=Ind Gender=Masc Number=Sing PronType=Art
2	Chile	chile	PROPN	PROPN___
3	con	con	ADP	ADP__AdpType=Prep
4	un	uno	DET	DET__Definite=Ind Gender=Masc Number=Sing PronType=Art
5	malestar	malestar	NOUN	NOUN__Gender=Masc Number=Sing
6	profundo	profundar	ADJ	ADJ__Gender=Masc Number=Sing
7	por	por	ADP	ADP__AdpType=Prep
8	tantas	tanto	DET	DET__Gender=Fem Number=Plur PronType=Ind
9	barreras	barrera	NOUN	NOUN__Gender=Fem Number=Plur
10	invisibles	invisible	ADJ	ADJ__Number=Plur

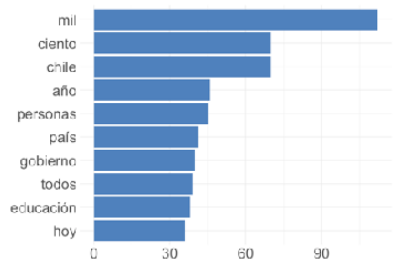
```
# ... with 17 more rows
```

~95%

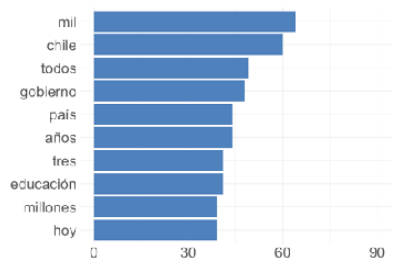
*accuracy* declarado por spaCy  
para el español

**retomemos el ejemplo inicial**

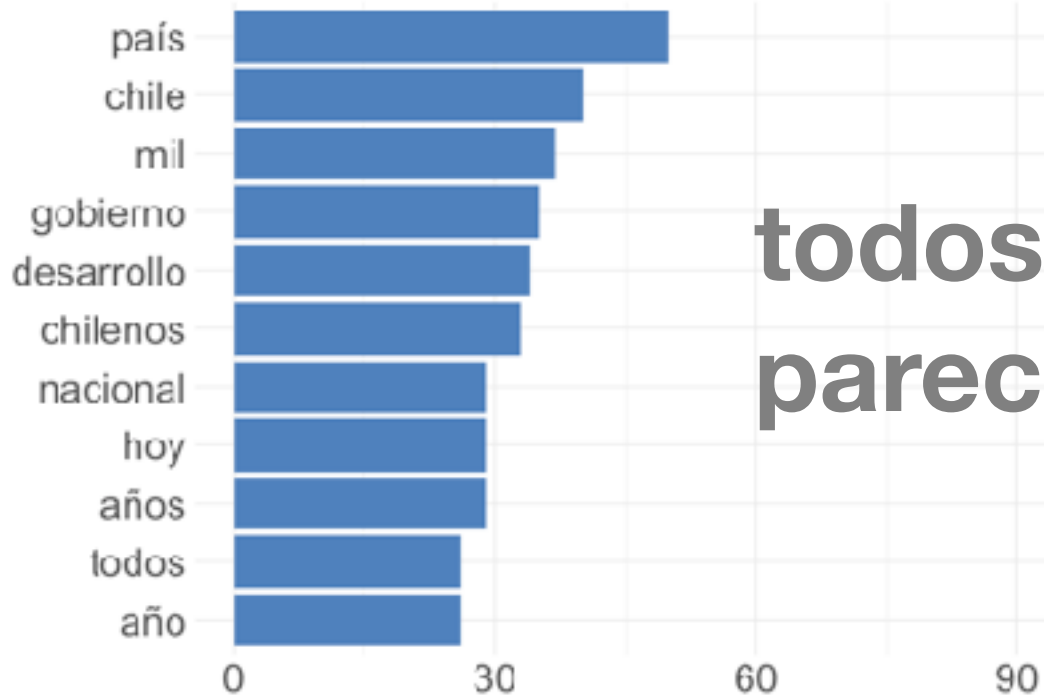
2017



2013



1989



**todos muy  
parecidos**

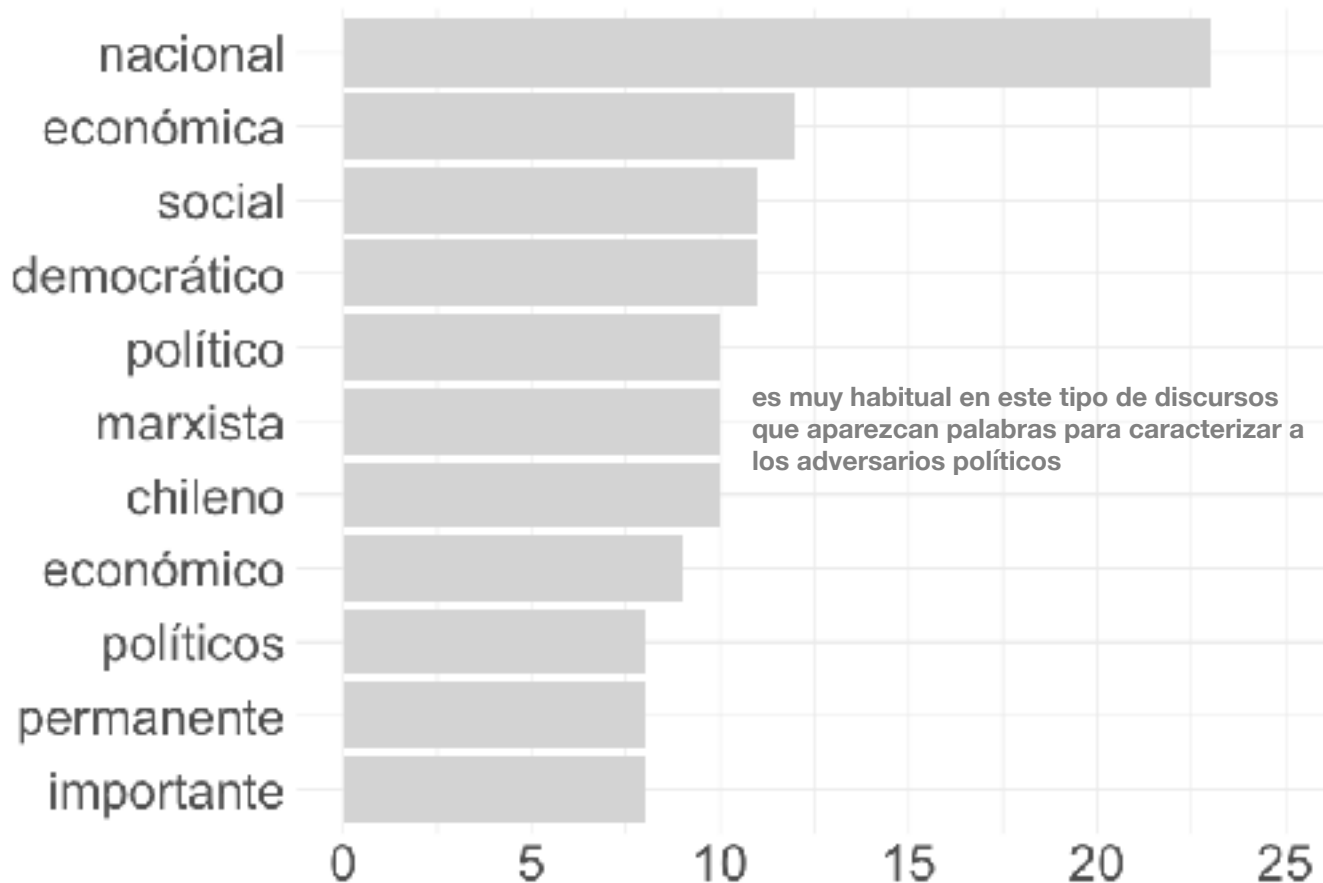
Para entender mejor lo que distingue a estos discursos podemos analizar una categoría gramatical como los adjetivos. Los hablantes tienen más opciones al elegirlos.

```
filter(upos == "ADJ")
```

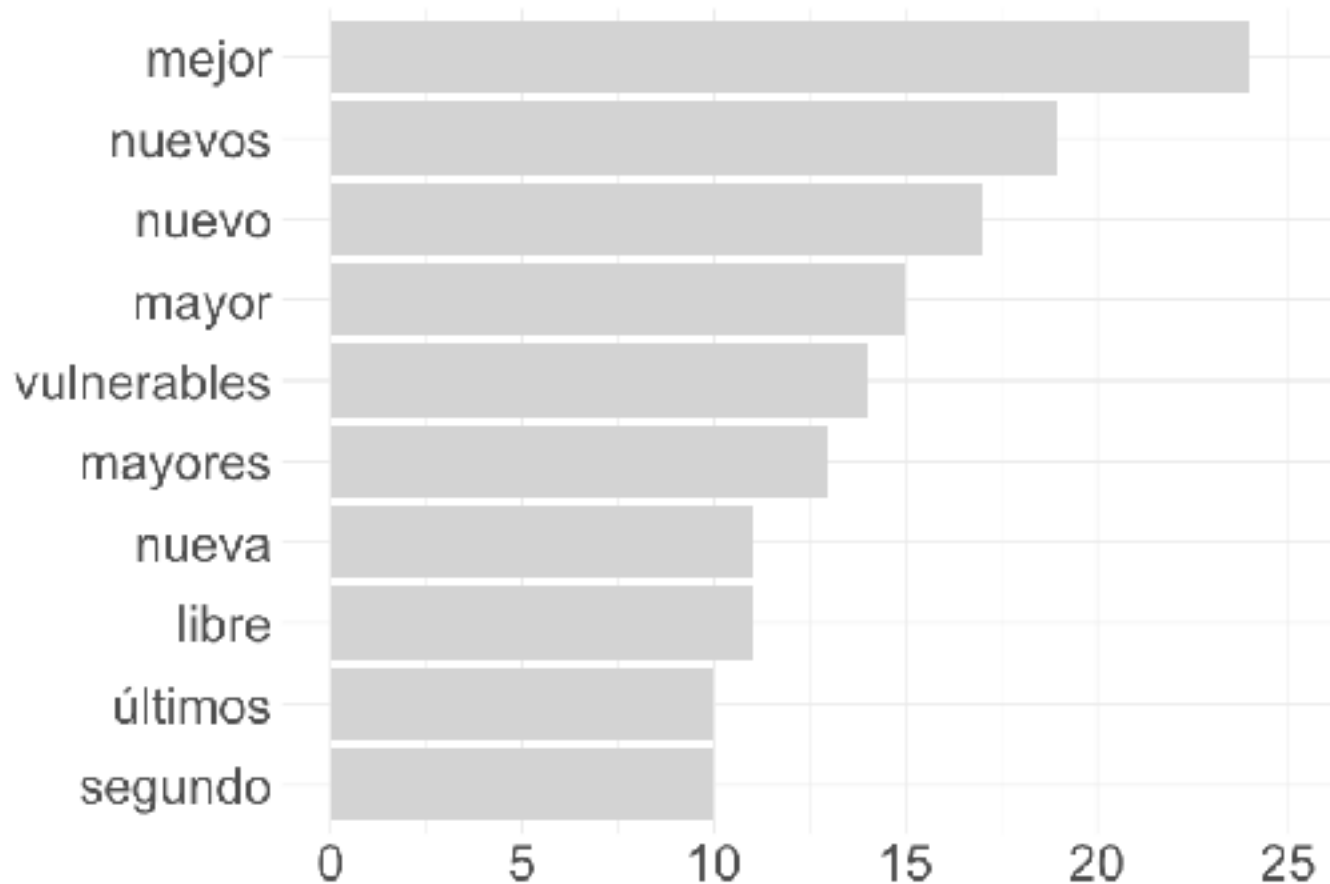


# 1989

## Pinochet

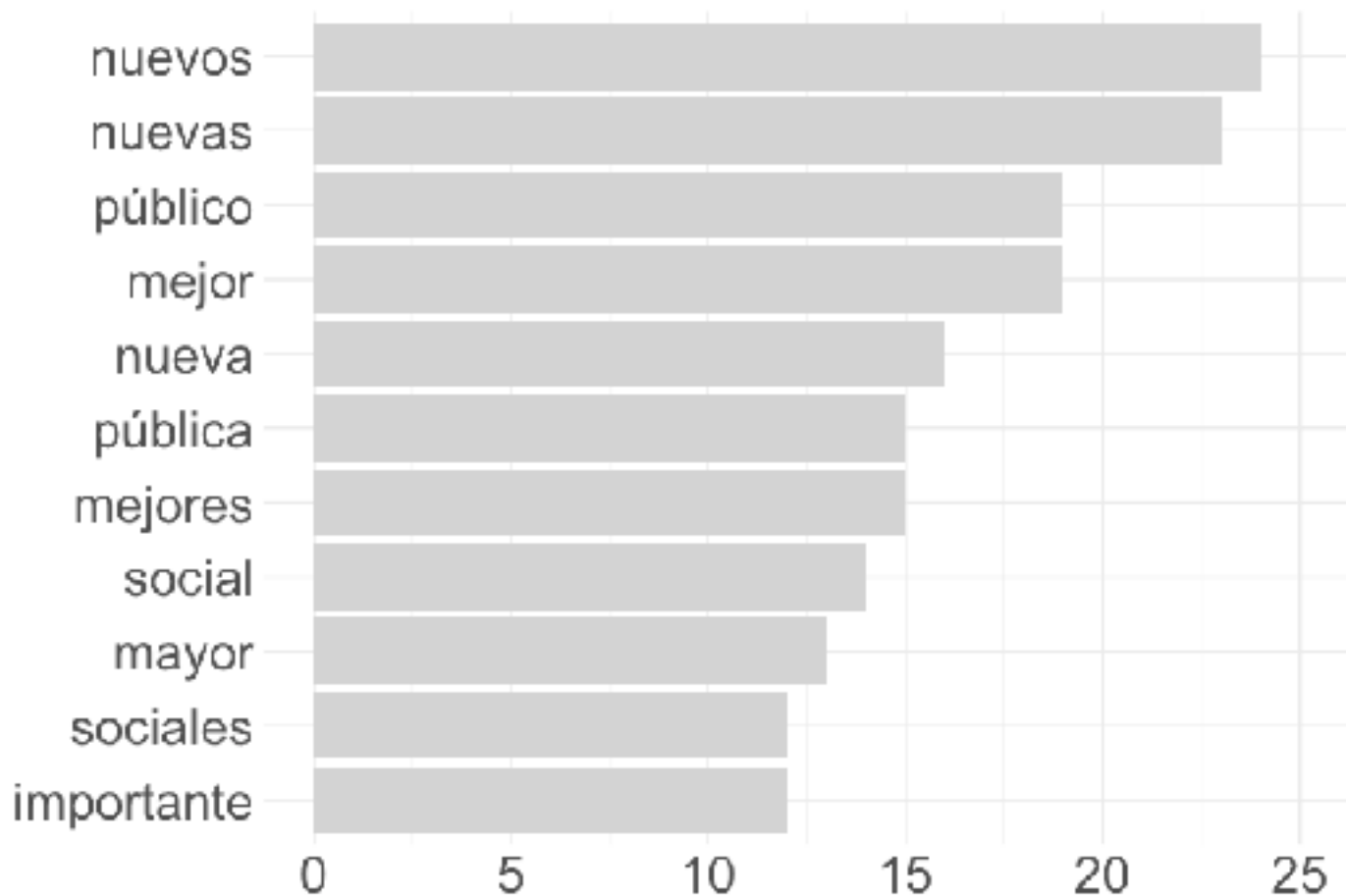


**2013**  
Piñera



# 2017

## Bachelet II



**Ahora sí podemos ver diferencias entre los mensajes. Los adjetivos utilizados en cada uno permiten caracterizar mejor su orientación política**

retomemos ahora esta idea

**“you shall know a word by the  
company it keeps”**

**(Firth 1957)**

**“you shall know a word by the  
company it keeps”**

**(Firth 1957)**

**análisis de  
concordancias**

**no es un análisis nuevo**

SACRORUM  
**BIBLIORUM**  
VULGATÆ EDITIONIS  
**CONCORDANTIÆ**  
HUGONIS CARDINALIS  
Ordinis Prædicatorum;  
AD RECOGNITIONEM  
JUSSU SIXTI V. PONT. MAX.  
BIBLIIS ADHIBITAM  
REGRNSITÆ, ATQUE EMENDATÆ.

*Primus à FRANCISCO LUGI Theologo, & Doctore Ardenasopolitano,  
postea sacre thesaurarius, ac licesseus eius, & postea  
V. D. HUGO PIERRE, Ordinis Sancti Benedicti,  
EDITIO NOVISSIMA, FRÆ CELEBIS CORRECTOR,  
in qua summo labore, ac diligentia singuli numeri et tractatus revisiti,  
atque emendati cum sacris Bibliis hanc donat integritate sunt.*



**VENETIIS. MDCCLIV.**

Apud Nicolaum Pezzana.

CUM PRIVILEGIO EXCELLENTISSIMI SENATUS.



# 1230

## Hugo de San Caro

publica una edición de la Biblia  
con un índice de concordancias:  
de cada término se ofrece el  
contexto en que aparece y su  
ubicación



SACRORUM  
**BIBLIORUM**  
VULGATÆ EDITIONIS  
**CONCORDANTIÆ**  
HUGONIS CARDINALIS  
Ordinis Prædicatorum;  
AD RECOGNITIONEM  
JUSSU SIXTI V. PONT. MAX.  
BIBLIIS ADHIBITAM  
REGRENSITÆ, ATQUE EMENDATÆ.

*Primus à FRANCISCO LUGI Theologo, & Doctore Astonsopolitano,  
postea vario loco repositus, ac lucubratu eius, & fidei  
U. D. HUGONIS PUGNETI, Ordinis Sancti Bernardi.*

*EDITIO NOVISSIMA FRÆ CELESTIS CORRECTOR,  
in qua summo labore, ac diligentia singuli numeri et tractatus revisiti,  
atque emendati cum sacris Bibliis hanc donat integritate sunt.*



VENETIIS. MDCCLIV.

Apud Nicolaum Pezzana.

CUM PRIVILEGIO EXCELLENTISSIMI SENATUS.

A. J. 1715



1230

Hugo de San Caro



(no estaba solo: un ejército de  
frailes lo ayudó a procesar el texto)

**En la década de 1960 se hizo por primera vez un análisis de este tipo de forma computacional**



## Roberto Busa



con el apoyo de IBM,  
inicia la elaboración de un  
índice de concordancias  
de la obra de Santo  
Tomás: el *Index  
Thomisticum*

<http://www.corpusthomisticum.org/it/index.age>

**¿lo hizo solo?**



**decenas de mujeres trabajaron en el proceso**

Ahora hay paquetes en R que nos  
permite hacer este tipo de análisis

**quanteda**  
Quantitative Analysis of Textual Data

y usando una sola función

kwi c()

(key word in context)

```
> kwic(corpus_mensajes, "migra", 5, "regex")
```

**ejemplo: queremos conocer cómo  
aparece el tema de la inmigración**



```
> kwic(corpus_mensajes, "migra", 5, "regex")
```

[pinochet_1987, 6569]	su tierra y evitando su	emigración	a la ciudad. Este
[aylwin_1990, 4565]	la Organización Internacional para las	Migraciones	, de 1987. Asimismo
[aylwin_1990, 13883]	costos sociales que acarrea la	migración	hacia las ciudades y el
[aylwin_1993, 1812]	la Organización Internacional para las	Migraciones	( OIM). Testimonio
[frei_ruiztagle_1999, 18979]	un grave efecto en la	migración	del campo a la ciudad
[bachelet_2006, 10139]	sufren los discapacitados, los	inmigrantes	, las minorías sexuales,
[bachelet_2017, 9792]	el influjo enriquecedor de la	migración	, muchos de los cuales
[bachelet_2017, 9829]	legislación a la nueva realidad	migratoria	. Nos medimos mejor,

**podemos ver el contexto de cada palabra y darnos cuenta que aluden a distintos tipos de inmigración**



**otro ejemplo: queremos saber  
cómo se ha abordado el tema de  
las comunicaciones satelitales**

```
kwic(corpus_mensajes, "sat(éle)l" , 5, "regex")
```

`kwic(corpus_mensajes, "sat(éle)l" , 5, "regex")`

[pinochet_1981, 451]	, pretende transformarnos en un	satélite	de la Unión Soviética.
[pinochet_1981, 4849]	una nueva Estación Terrestre de	Satélites	, que unirá la XI
[pinochet_1981, 4890]	, permitirá el uso de	satélites	para comunicaciones domésticas y
[pinochet_1983, 381]	transformar al país en un	satélite	del imperialismo soviético. La
[pinochet_1985, 1650]	convertir al país en un	satélite	más de la Unión Soviética
[pinochet_1988, 7518]	extraviadas, a base de	satélites	, avance que es el
[pinochet_1989, 1033]	nuestra Patria en un nuevo	satélite	de la Unión Soviética.
[pinochet_1989, 8397]	paramilitares, adiestrados en países	satélites	de la Unión Soviética,
[pinera_2012, 1579]	, que incluye 78 teléfonos	satelitales	en lugares estratégicos,
[pinera_2012, 6623]	Chile cuenta actualmente con el	satélite	Fasat-Charlie, que permite la
[pinera_2012, 6652]	trabajando en una política nacional	satelital	y en un proyecto de
[pinera_2012, 6658]	y en un proyecto de	satélite	de telecomunicaciones,
[bachelet_2017, 15564]	más recursos, con tecnología	satelital	y un plan estratégico para

**Hay casos en los que se utiliza con un sentido muy distinto. Si no hubiésemos mirado el contexto y solo nos hubiésemos quedado con las frecuencias, podríamos haber malinterpretado la alta frecuencia de este término en los mensajes de Pinochet**

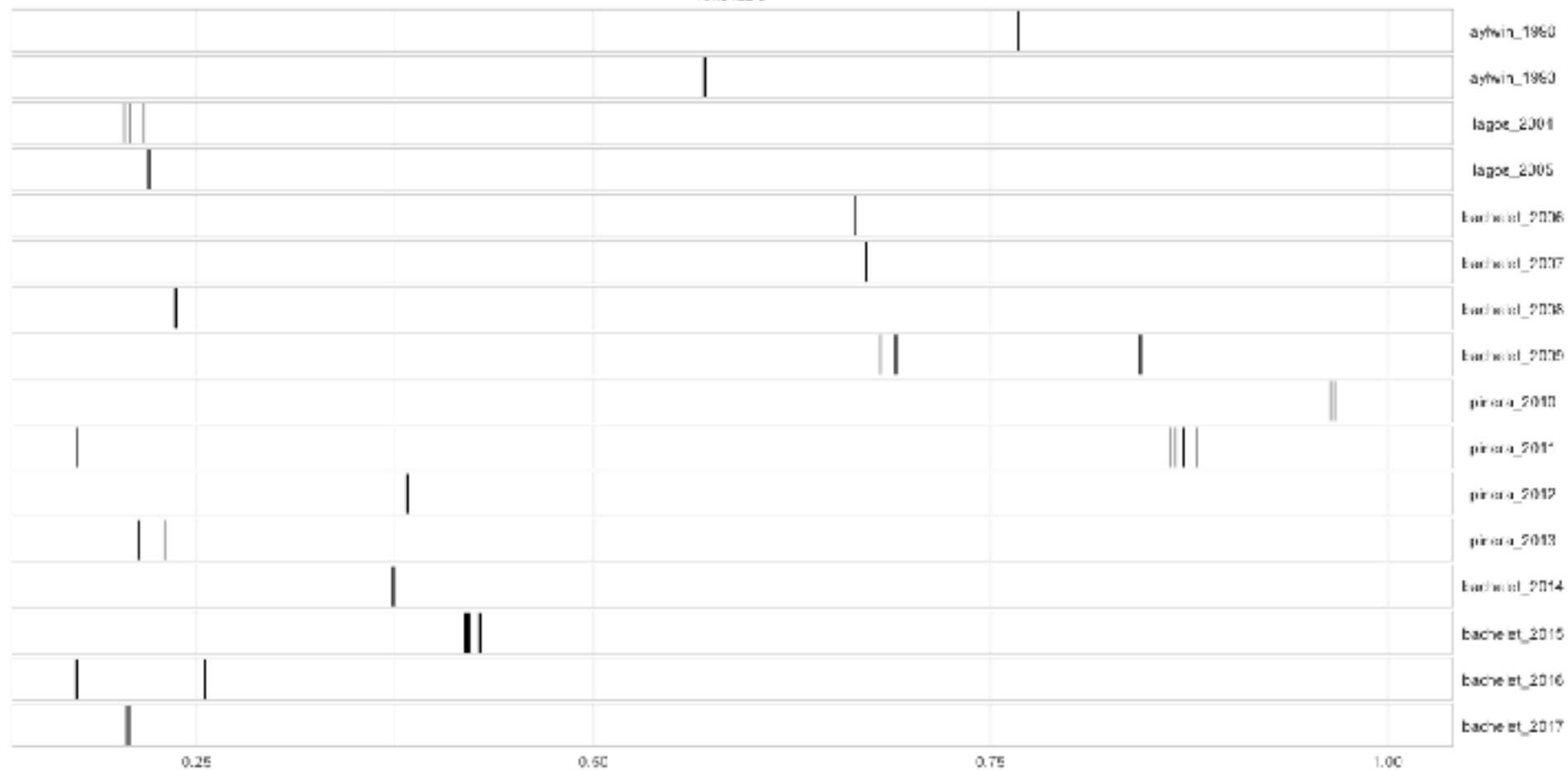
**¿y la distribución?**

**También resulta importante conocer en qué parte del texto aparece un determinado término. En algunos casos, que aparezcan antes puede significar que son considerados más importantes.**

```
corpus_mensajes %>%  
  kwic("renovable", valuetype = "regex") %>%  
  textplot_xray()
```

# Dispersión léxica

renovable

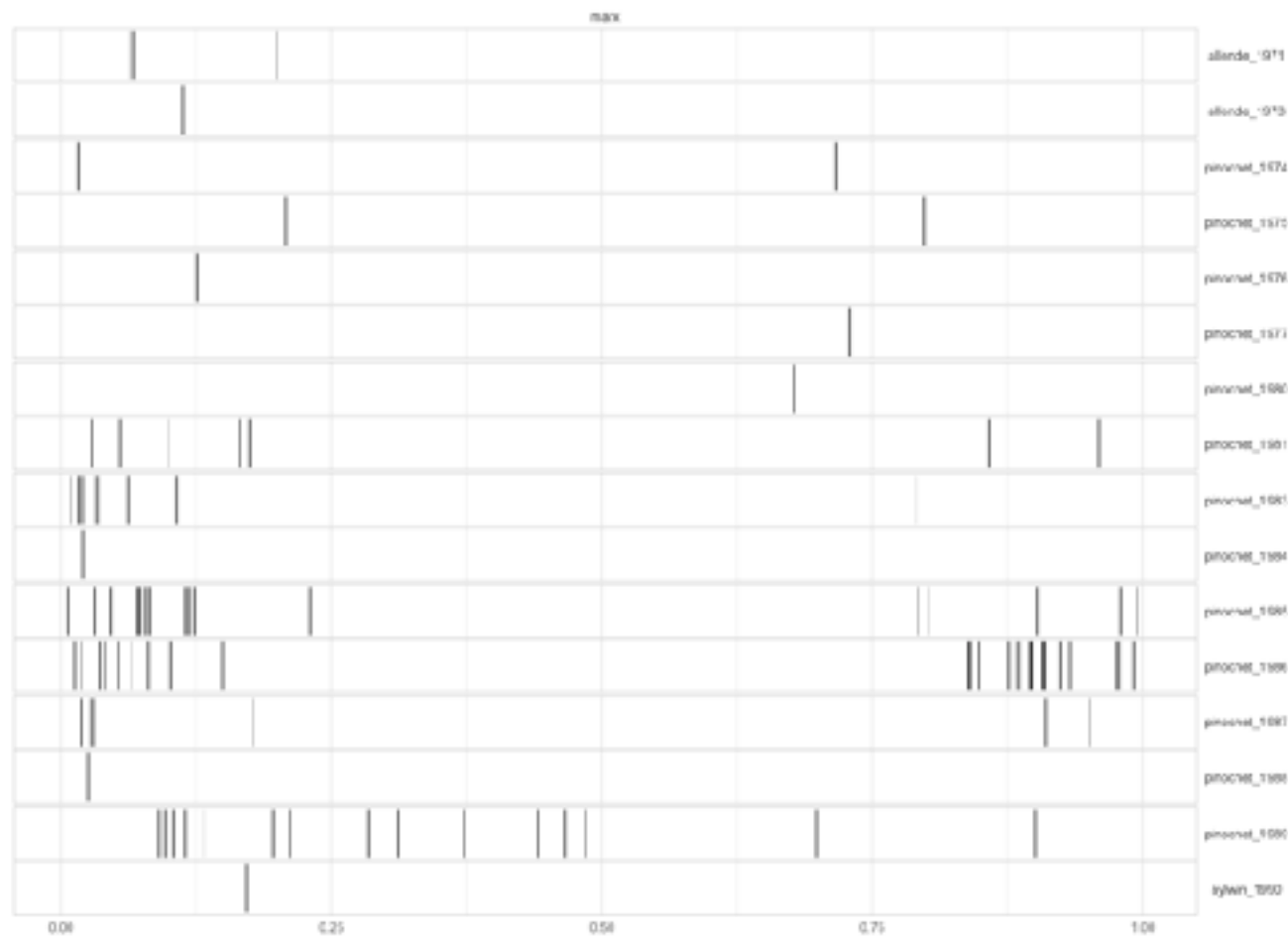




**Al inicio de la presentación vimos que “marxista” era un término que aparecía con frecuencia en los discursos de Pinochet. ¿En qué parte lo utiliza?**

kwic("marx")

Diapers en l'exica



**¿y las comparaciones  
entre discursos?**

**keyness**

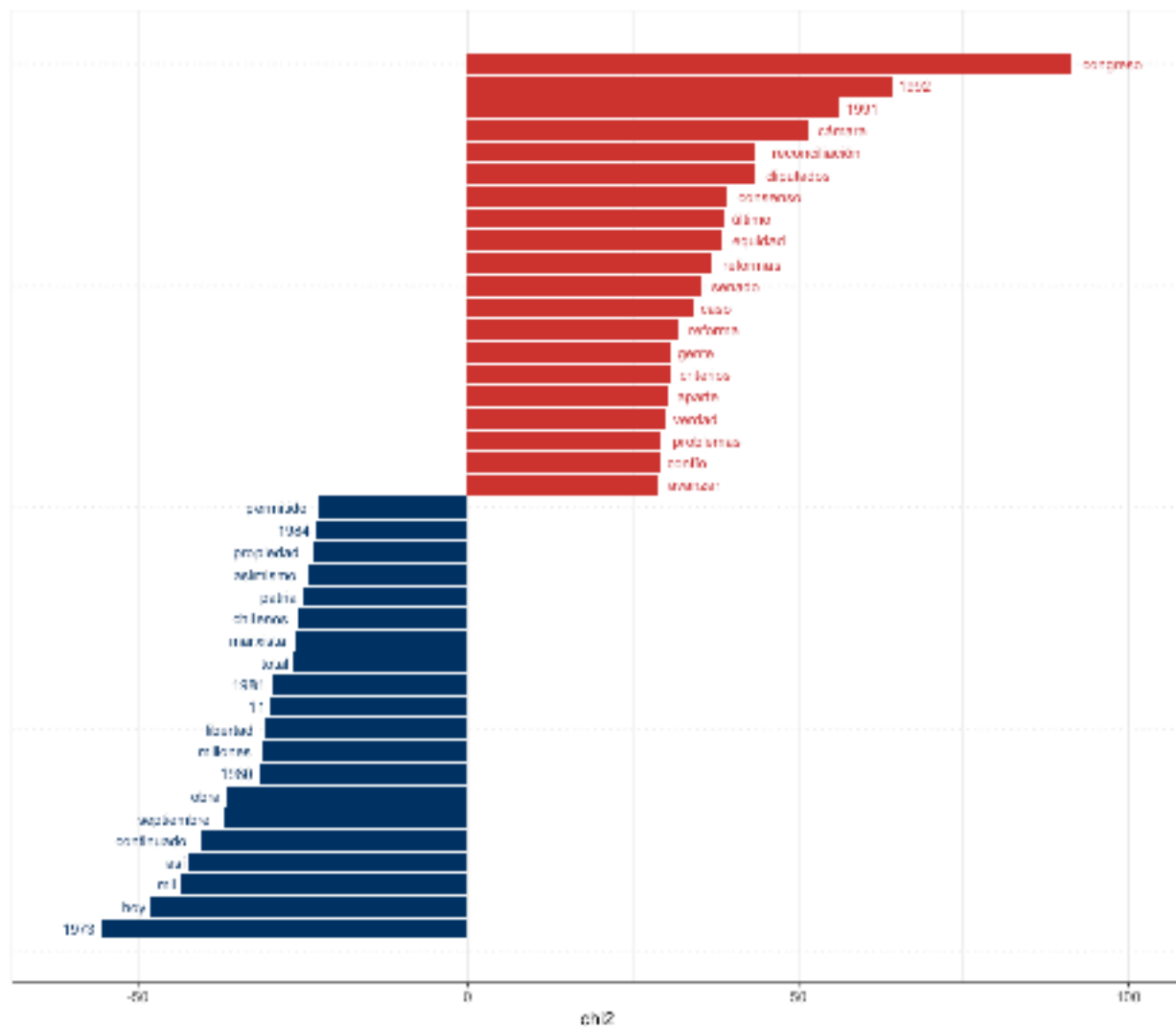
**Pinochet / Aylwin**

```
dict_trans <- corpus_subset(corpus_mensajes, Presidente %in% c("Augusto  
Pinochet", "Patricio Aylwin"))
```

```
dict_trans_dfm <- dfm(dict_trans, groups = "Presidente", remove =  
stopwords("es"), remove_punct = TRUE)
```

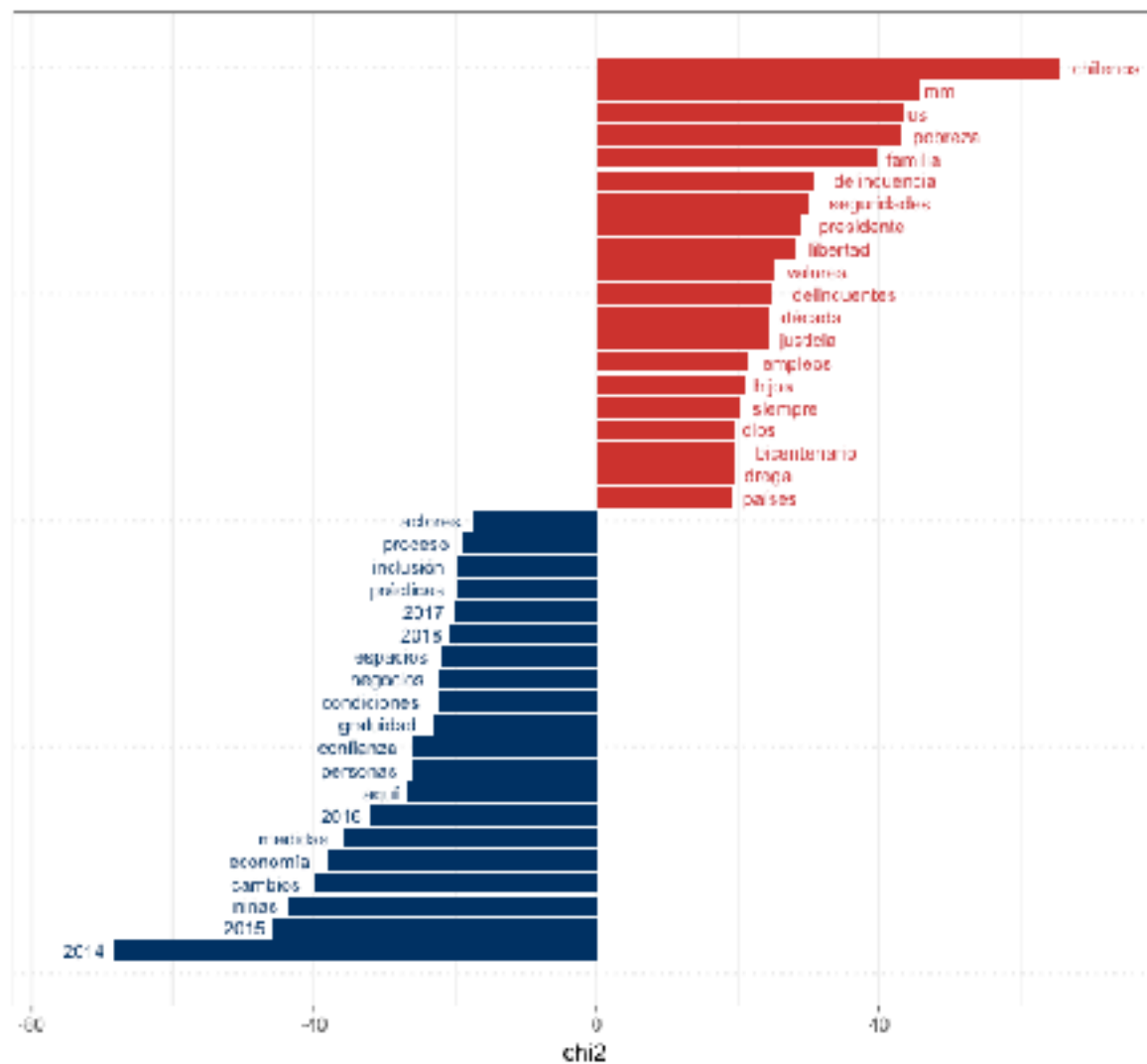
```
result_keyness <- textstat_keyness(dict_trans_dfm, target = "Patricio  
Aylwin")
```

```
textplot_keyness(result_keyness)
```



En este caso era interesante no remover los números

**Piñera / Bachelet II**



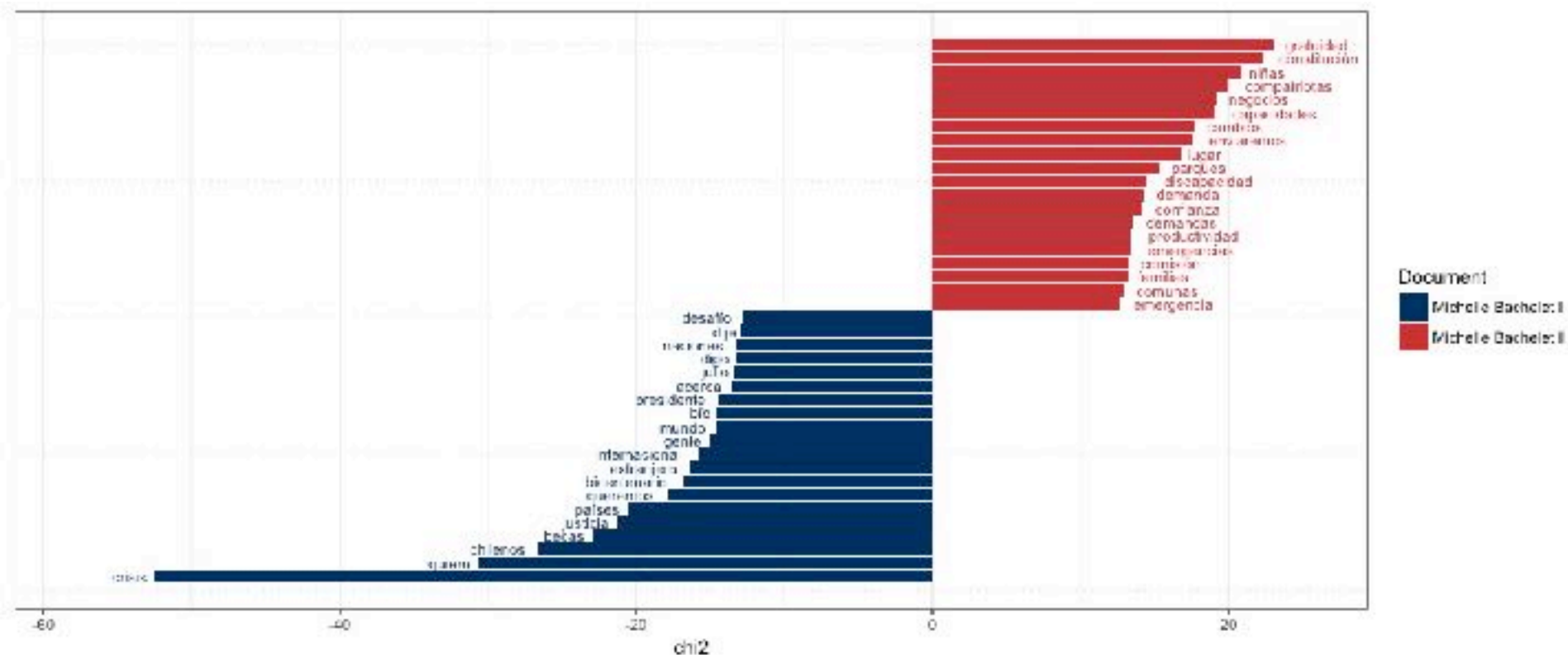
#### Document

- Michelle Bachelet II
- Sebastián Piñera

Los gráficos son editables con ggplot2. Así se ven por defecto.



**Bachelet I / Bachelet II**



**¿Qué me interesa hacer con todo esto?**

**Un sitio web que permita a sus  
usuarios explorar los discursos**

`mensajespresidenciales.cl`

**¡Estará listo en mayo!**



```
library(dplyr)
```

```
rladies_global %>%  
  filter(city == 'Santiago')
```

# Análisis de textos con R

Búsqueda de patrones lingüísticos  
en discursos políticos

Riva Quiroga



@rivaquioga